Machine Learning for Static and Single-Event Dynamic Complex Network Analysis

Nakis, Nikolaos

Publication date:
2023

Document Version
Publisher's PDF, also known as Version of record

Ph.D. Thesis
September 2023

DTU Compute
Department of Applied Mathematics and Computer Science

Machine Learning for Static and Single-Event Dynamic Complex Network Analysis

Nikolaos Nakis

DTU

Ph.D. Degree program

Technical University of Denmark (DTU),
Department of Applied Mathematics and Computer Science

Supervisor: Prof. Morten Mørup (DTU)
Co-supervisor: Prof. Sune Lehmann (DTU)

DTU Compute
Department of Applied Mathematics and Computer Science
Technical University of Denmark
Building 321
2800 Kongens Lyngby, Denmark

Summary (English)

Networks are prevalent data structures that naturally express complex systems. They emerge across a multitude of scientific domains, including physics, sociology, the science of science, biology, neuroscience, and more. In these disciplines, networks illustrate diverse interactions and systems: spin glasses in physics, social connections in sociology, academic collaborations, protein-to-protein interactions in biology, and both structural and functional brain connectivity in neuroscience, to name a few. Due to their complexity and inherently high-dimensional discrete nature, accurately characterizing network structures is both non-trivial and challenging. In recent years, Graph Representation Learning (GRL) has achieved remarkable success in the study of networks, establishing itself as the leading method for network analysis. In general, GRL aims to create a function that can successfully map a network to a low-dimensional latent space through a learning process. Such a mapping defines representations that can be very useful for conducting various downstream tasks, and importantly for helping us to further our understanding of complex networks and their underlying structures.

The primary objective of this thesis is to develop novel algorithmic approaches for Graph Representation Learning of static and single-event dynamic networks. In such a direction, we focus on the family of Latent Space Models, and more specifically on the Latent Distance Model which naturally conveys import network characteristics such as homophily, transitivity, and the balance theory. Furthermore, this thesis aims to create structural-aware network representations, which lead to hierarchical expressions of network structure, community characterization, the identification of extreme profiles in networks, and impact dynamics quantification in temporal networks. Crucially, the methods presented are designed to define unified learning processes, eliminating the need for heuristics and multi-stage processes like post-processing steps. Our aim is to delve into a journey towards unified network embeddings that are both comprehensive and powerful, capable of characterizing network structures and adeptly handling the diverse tasks that graph analysis offers.

Summary (Danish)

Netværk er almindelige datastrukturer, der naturligt udtrykker komplekse systemer. De opstår på tværs af mange videnskabelige domæner, herunder fysik, sociologi, videnskabens videnskab, biologi, neurovidenskab og mere. I disse discipliner illustrerer netværk forskellige interaktioner og systemer: spin-glas i fysik, sociale forbindelser i sociologi, akademisk samarbejde, protein-til-protein interaktioner i biologi, samt både strukturel og funktionel hjerneforbindelse i neurovidenskab, for blot at nævne nogle få. På grund af deres kompleksitet og iboende høj-dimensionale diskrete natur er nøjagtig karakterisering af netværksstrukturer både ikke-trivielt og udfordrende. I de seneste år har Graf Representation Læring (GRL) opnået bemærkelsesværdig succes i studiet af netværk, og har etableret sig som den førende metode til netværksanalyse. Generelt sigter GRL mod at skabe en funktion, der med succes kan kortlægge et netværk til et lav-dimensionalt latent rum gennem en læringsproces. En sådan kortlægning definerer repræsentationer, der kan være meget nyttige til at udføre forskellige efterfølgende opgaver, og vigtigt for at hjælpe os med yderligere at forstå komplekse netværk og deres underliggende strukturer.

Hovedformålet med denne afhandling er at udvikle nye algoritmiske tilgange til Graf Representation Læring af statiske og enkeltbegivenheds dynamiske netværk. I denne retning fokuserer vi på familien af Latente Rummodeller, og mere specifikt på den Latente Afstandsmodel, som naturligt formidler vigtige netværksegenskaber såsom homofili, transitivitet og balance teorien. Yderligere sigter denne afhandling mod at skabe strukturbevidste netværksrepræsentationer, hvilket fører til hierarkiske udtryk af netværksstruktur, fællesskabskarakterisering, identifikation af ekstreme profiler i netværk og kvantificering af påvirkningsdynamik i tidsmæssige netværk. Afgørende er de præsenterede metoder designet til at definere ensartede læringsprocesser, hvilket eliminerer behovet for heuristikker og flertrinsprocesser som efterbehandlings trin. Vores mål er at dykke ned i en rejse mod ensartede netværksindlejringer, der er både omfattende og kraftfulde, i stand til at karakterisere netværksstrukturer og dygtigt håndtere de forskelligartede opgaver, som grafanalyse tilbyder.

Preface

This Ph.D. thesis, entitled *Machine Learning for Static and Single-Event Dynamic Complex Network Analysis*, was prepared in the Section for Cognitive Systems (CogSys) within the Department of Applied Mathematics and Computer Science at the Technical University of Denmark (DTU). It is submitted in partial fulfillment of the requirements for obtaining a Ph.D. in Applied Mathematics and Computer Science from DTU.

The Ph.D. project was supervised by Prof. Morten Mørup and co-supervised by Prof. Sune Lehmann while it was generously financed by the Independent Research Fund Denmark [grant number: 0136-00315B]. The Ph.D. project was carried out at DTU during the period September 2020 - September 2023, except for a five-month external stay at the University of Umeå under the supervision of Prof. Martin Rosvall, and a three-month external stay at Yale University under the supervision of Sterling Prof. Nicholas Christakis.

The thesis comprises five research papers focused on the topic of Graph Representation Learning.

Kongens Lyngby, 10$^{\text{th}}$ September 2023

Nikolaos Nakis

Acknowledgements

First and foremost, I would like to thank my supervisor, Prof. Morten Mørup. I consider myself extremely fortunate to have had such a talented and unique researcher guide me through my Ph.D. studies and act as my mentor. From my very first day on this journey, Morten has been a constant pillar of support, standing by me through the 'good' and 'bad' surprises our research presented. I will forever be grateful for all the things I have learned alongside him, which will stay with me for the rest of my career. It's largely thanks to him that I decided to pursue a career in research.

I am eternally grateful to Dr. Abdulkadir Çelikkanat. Kadir has been a consistent source of inspiration throughout my Ph.D. journey. His natural talent for research, his exceptional work ethic, and his kind personality have all guided me toward becoming a better researcher and, more broadly, a better person. I owe him immense gratitude for all the support and guidance he has provided these past years.

I would like to thank my co-supervisor, Prof. Sune Lehman. Sune is not only one of the most outstanding researchers I have had the pleasure of working with but also one of the coolest. Over the past few years, he has consistently been a source of inspiration, brimming with amazing and endless research ideas. His guidance has significantly broadened my research horizons and shaped my way of thinking.

Throughout this journey, I was fortunate to encounter exceptional individuals and researchers who I now proudly call friends. Tremendous thanks to Davide and Agatha; spending time with you both has helped me navigate the lows and celebrate the highs of these past years, you have been my Polish-Italian hybrid family in Copenhagen. In the same spirit, my deep appreciation goes to Peter, Germans, Mohammed (Mo), George, Giorgio, Lasse, Louis, Silvia, and Rui. Thank you all for everything.

I have had the pleasure to be a part of the CogSys research family. I would like to thank all of my colleagues at DTU for becoming a part of my daily routine these past years.

A special thanks to my dearest friend George our friendship united us also in our academic choices and since the day we met in Denmark our life paths have been very similar. Thank you for all the discussions, laughs, and frustrations we shared during

our Ph.D. journeys; they truly helped me reach where I am today.

I cannot find the words to express my gratitude to you, Anastasia. From the first day we met in Denmark, you have consistently been there for me, guiding and inspiring me with your passion for research and life. I owe part of who I've become today to you, and I'm deeply thankful for all the help you've generously provided, enabling me to reach this milestone.

Lastly, I would like to thank my family and friends. Especially, to my parents Grigoris and Aristi, I would not be here without you. Thank you for everything.

List of publications

Contributions included in the thesis:

1. N. Nakis, A. Çelikkanat, S. Lehmann and M. Mørup, "A Hierarchical Block Distance Model for Ultra Low-Dimensional Graph Representations," in *IEEE Transactions on Knowledge and Data Engineering*, https://doi.org/10.1109/TKDE.2023.3304344, 2023.

2. N. Nakis, A. Çelikkanat, and M. Mørup. "HM-LDM: A Hybrid-Membership Latent Distance Model". In: *Complex Networks and Their Applications XI*. COMPLEX NETWORKS 2016 2022. Studies in Computational Intelligence, vol 1077. Springer, Cham. https://doi.org/10.1007/978-3-031-21127-0_29

3. N. Nakis, A. Çelikkanat, L. Boucherie, C. Djurhuus, F. Burmester, D. M. Holmelund, M. Frolcová, and M. Mørup. "Characterizing Polarization in Social Networks using the Signed Relational Latent Distance Model". *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, PMLR 206:11489-11505, 2023.

4. N. Nakis, A. Çelikkanat, and M. Mørup. "A Hybrid Membership Latent Distance Model for Unsigned and Signed Integer Weighted Networks". In: *Advances in Complex Systems*. https://doi.org/10.1142/S0219525923400027, 2023.

5. N. Nakis, A. Çelikkanat, and M. Mørup. "Time to Cite: Modeling Citation Networks using the Dynamic Impact Single-Event Embedding Model". *Preprint*, 2023.

Contributions not included in the thesis:

6. A. Çelikkanat, N. Nakis, and M. Mørup, "Piecewise-velocity model for learning continuous-time dynamic node representations," in *Proceedings of the First Learning on Graphs Conference* (B. Rieck and R. Pascanu, eds.), vol. 198 of Proceedings of Machine Learning Research, pp. 36:1–36:21, PMLR, 09–12 Dec 2022.

7. A. Çelikkanat, N. Nakis, and M. Mørup. Continuous-time Graph Representation with Sequential Survival Process. *Preprint*, 2023.

Work related to these papers was presented at the following conferences:

(*) indicates the presenting author

1. N. Nakis*, A. Çelikkanat, and M. Mørup. "HM-LDM: A Hybrid-Membership Latent Distance Model". *The 11th International Conference on Complex Networks and their Applications*, Palermo (Italy), Nov 2022, Oral Presentation.

2. A. Çelikkanat*, N. Nakis, and M. Mørup. Piecewise-Velocity Model for Learning Continuous-time Dynamic Node Representations". *The 1st Learning on Graphs Conference*, Online, Dec 2022, Poster Presentation.

3. N. Nakis, A. Çelikkanat, L. Boucherie, C. Djurhuus*, F. Burmester*, D. M. Holmelund*, M. Frolcová, and M. Mørup. "Characterizing Polarization in Social Networks using the Signed Relational Latent Distance Model". *The 25th International Conference on Artificial Intelligence and Statistics*, Valencia (Spain), March 2023, Poster Presentation.

4. N. Nakis*, A. Çelikkanat, L. Boucherie, and M. Mørup. "Characterizing Polarization in Social Networks using Archetypal Analysis". *The 9th International Conference on Computational Social Science*, Copenhagen (Denmark), Jul 2023, Oral Presentation.

5. A. Çelikkanat*, N. Nakis, and M. Mørup. "Piecewise-Velocity Model for Learning Continuous-time Dynamic Node Representations". *The 9th International Conference on Computational Social Science*, Copenhagen (Denmark), Jul 2023, Oral Presentation.

List of Symbols

Symbol	Description
$\mathcal{G}$	Graph
$\mathcal{V}$	Vertex set
$\mathcal{E}$	Edge set
$\mathcal{E}^+$	Positive edge set
$\mathcal{E}^-$	Negative edge set
N	Number of nodes
D	Dimension size
$\gamma_i, \beta_i, \psi_i, \alpha_i$	Bias terms of node i
$\mathbf{w}_i, \mathbf{z}_i$	Latent embeddings for node i
λ_{ij}	Poisson rate (intensity) of node pair (i, j)
λ_{ij}^+	Positive interaction Poisson rate (intensity) of node pair (i, j) of the Skellam distribution
λ_{ij}^-	Negative interaction Poisson rate (intensity) of node pair (i, j) of the Skellam distribution
$\mathcal{I}_{\lvert y\rvert}$	Modified Bessel function of the first kind and order $\lvert y\rvert$
δ	Simplex side length with $\delta \in \mathbb{R}_+$
p	Power of the ℓ_2 norm with $p \in \{1, 2\}$
Δ^D	The standard $D-$simplex
$\mathbf{\Lambda}$	Eigenmodel non-negative relational matrix
$\mathbf{A}$	The matrix containing the archetypes (extreme points of the convex hull)
$\boldsymbol{\mu}_i$	cluster centroid vector
μ	mean value of a distribution
σ	standard deviation of a distribution
T	single-event network timeline

Contents

Part I

Introduction and methods

CHAPTER 1

Introduction

1.1 Networks

Networks are widespread data structures and represent the most natural means of expressing complex systems. They appear across various scientific domains, encompassing fields such as physics, sociology, science of science, biology, and more. Within these disciplines, networks are used to describe a multitude of interactions and systems, such as spin glasses in physics, friendship interactions in sociology, scholarly collaborations in academia, protein-to-protein interactions in biology, and structural and functional brain connectivity in neuroscience, among many others [1]. Given their complexity and high-dimensional discrete nature, accurately characterizing the structure of networks is regarded as a non-trivial and challenging task. Scientists employ various graph analysis tools to examine these networks, seeking to gain insights into their underlying structures. These tools are used for several downstream tasks, including link/relation prediction [2], node classification and clustering [3,4], and community detection [5, 6]. Moreover, the importance of network analysis extends beyond scientific research, influencing practical applications in industries like telecommunications, transportation, healthcare, and finance. Whether optimizing routes in a transportation network, understanding the spread of diseases within a population, or detecting fraudulent activities within a financial system, network analysis plays a pivotal role. The methodologies and techniques developed in network analysis continue to evolve, pushing the boundaries of our understanding and application of complex systems. The synergy between theoretical development and practical application ensures that network analysis remains an integral and dynamic field of study, connecting diverse domains and contributing to advancements across a broad spectrum of disciplines.

1.1.1 Network science

Towards advancing our understanding of networks, network science emerged. A multidisciplinary field that focuses on the study of complex networks, searching characterizations for their structure, behavior, evolution, and function. It incorporates concepts and methodologies from areas such as physics, mathematics, computer science, biology, sociology, and economics, allowing for a comprehensive analysis and understanding of various kinds of networks [7]. Structural analysis for complex sys-

tems focuses on the examination of the node and edge properties of a given network, including statistics such as degree distribution, clustering coefficients, and community structures. Seminal work in this area includes the studies of small-world networks [8] and scale-free networks [9]. Temporal network analysis under dynamic processes aims to further our understanding of how networks evolve and change over time, and how processes such as information spreading, disease propagation, and social influence operate on such structures. Classic models like the Erdős–Rényi model [10] modeling the evolution of a random network and the SIR model [11] for disease spreading were the pioneering works. Analysis of multilayer complex networks investigates networks that have multiple types or layers of connections. This area explores how different layers interact with each other and contribute to the overall behavior of the network. Works like the model of interconnected networks [12] have deepened our understanding of these complex systems. Important directions also include network visualization and data mining where researchers utilize computational and visualization tools to analyze large-scale network data. This includes discovering patterns, anomalies, and community structures within big data sets. Identifying influential spreaders in complex networks [13] and designing community detection algorithms [14] are prominent examples of data mining in network science. Being a multidisciplinary field, network science defines multiple applications across various fields (for a comprehensive overview please see [15]). Through a combination of mathematical modeling, computational analysis, and empirical study, network science continues to offer profound insights into the structures and dynamics that underlie diverse systems.

1.1.2 Classical methods for network analysis

The very first approaches to studying networks focused on node properties and node-level statistics. These included various centrality measures [16], like node degree, eigenvector centrality [17], and the clustering coefficient [8], to name a few. Centrality measures focus on expressing different formulations for the importance of nodes in a network, able to capture various properties. Apart from centrality measures, multiple metrics of similarity in terms of the node neighborhood overlap have also been proposed and extensively studied. These include local overlap measures [18] like the Jaccard overlap, the Sorensen index, and the Adamic-Adar index which express different functions over each node's local neighborhood and the common neighbors that two nodes share. Such metrics define node similarity while accounting for the node degree biases with variations on the expression of importance that each common neighbor provides to the metric. For example, the Adamic-Adar index gives higher importance to connections with lower-degree nodes, as they are regarded as more informative than connections with high-degree nodes. Similarly, various global overlap measures have also been proposed which also take into account the global network structure (rather than only the local neighborhood). Popular choices include the Katz Index [19] which counts over the number of paths of all lengths between a pair of nodes, the Leicht, Holme, and Newman similarity [20] correcting over the Katz

Index to account for degree biases by normalizing with the number of expected paths between two nodes, and various random walk methods such as the PageRank [21] algorithm, expressing stationary probabilities that a random walk starting at node i has to visit node j at some point. Local and global measures express multiple important characteristics of networks and often provide competitive performance in the downstream tasks even against models with advanced learning procedures [22] but express limitations due to their heuristic nature.

The early algorithmic attempts towards obtaining graph representations relied on spectral-decomposition approaches under dimensionality reduction frameworks defining an approximative expression of the Laplacian or adjacency matrices [23,24]. Classical examples include the Isomap algorithm [25] which uses Multidimensional Scaling (MDS) [26] in order to translate the k-nearest neighbors-based geodesic distances between nodes into a lower dimensional Euclidean space. Another type of well-known methods are the Laplacian Eigenmaps [27–29], where node embeddings are defined as the k-smallest eigenvectors of the normalized graph Laplacian. Laplacian matrices have found lots of applications in graph analysis due to the rich cut information [23] they carry. Matrix factorization approaches have been studied extensively and in-depth due to their simplicity, with a lot of linear and non-linear variants [30–32]. In contrast to their many advantages, these methods can be expensive when analyzing large networks due to the computational cost that the matrix decomposition enforces, a study trying to solve the scalability issues via graph partition and parallel computation was proposed in [33].

1.2 Graph Representation Learning

Towards the understanding of networks, Graph Representation Learning (GRL) [34, 35] has found incredible success in the past years, regarded as the foremost method for network analysis. GRL has been so popular since it is composed of approaches outperforming significantly the prior classical methods in the downstream tasks. Classic methods usually rely on graph kernels and graph statistics including various graph centrality measures [34]. Unlike GRL, conventional algorithms exhibit restricted flexibility and capacity as they employ node and graph-level statistics, requiring meticulous heuristic design and often resulting in high time and space complexity [34]. The main goal of GRL is to construct a function that defines a mapping of the network into a low-dimensional (usually Euclidean) latent space through a learning process. Specifically, such a projection has to translate node, edge or even graph similarity of network(s) into similarity in the latent space, i.e., by positioning related nodes, edges, or graph representations close in proximity in the latent space [36]. The main focus of GRL lies in learning continuous vector representations for individual nodes, edges, or graphs in the graph-defining embeddings. Node representations can be used for tasks like node classification, link prediction, and clustering. Influential methods in this area include DeepWalk [22], Node2Vec [4], and LINE [37]. Edge embeddings

are most often used to predict or infer missing links within a network. Techniques such as GraphSAGE [38] and SEAL [39] have contributed significantly to this aspect of GRL. Graph embeddings focus on learning a comprehensive representation of the entire graph, which can be employed in graph classification or similarity computation and is considered one of the most complex GRL areas. Graph Kernels [40] and Graph Neural Networks (GNNs) [41] have shown great success in this domain. A lot of attention has been given to Graph Neural Networks (GNNs) which define deep learning models specifically designed to operate on graph data. Convolutional Graph Neural Networks (GCNs) [42] and Graph Attention Networks (GATs) [43] are prominent examples. Importantly, GNNs focus on combining structural information with node features, labels, and other metadata to enrich the learning process. This has been shown to improve results in tasks like node classification, as seen in methods like HAN [44]. A prominent requirement in GRL are scalability and efficiency aspects. In particular, a major aim is the developing of algorithms and techniques that can scale to large graphs while maintaining computational efficiency which we will also address in this thesis. Techniques like GraphSAGE [38] and FastGCN [45] address these challenges. Lastly, GRL aims to incorporate higher-order proximity unlike most traditional approaches; accounting for both direct and indirect relationships among nodes to provide richer and more nuanced embeddings [46].

1.2.1 Matrix decomposition methods

A notable category for GRL relies on matrix decomposition techniques [35, 46–48] where node representations are obtained by the decomposition of a target matrix, constructed in such a way as to convey nodal proximity information, and can potentially include both first and higher-order adjacency information [46, 49]. The core idea lies in the assumption that the defined target matrix can be represented by a small number of latent factors that constitute the node embeddings. The main drawback of such methods is their space and time complexities since they usually lead to quadratic dependencies with the number of nodes in the graph. Recent studies aimed to address such computational challenges through techniques like matrix sparsification tools, hierarchical representations, or fast hashing schemes [47, 50–53].

1.2.2 Random walk methods

Pioneering GRL approaches drew inspiration from Natural Language Processing (NLP) [54] and employed random walks to generate node sequences analogous to sentences in NLP [4, 22, 37, 55]. Specifically, these works leveraged the Skip-Gram algorithm [56], to acquire node representations [4, 22, 37, 55, 57] by optimizing the co-occurrence probability for node pairs based on their distances obtained through the walks. The initial representatives of the random walk-based methods are DeepWalk [22] and Node2Vec [4]. As an extension to the DeepWalk procedure, Node2Vec introduced a global walk bias which allowed the use of both Breadth-First Sam-

pling and Depth-First Sampling during training in order to learn both local and global graph structures. Random walk methods are actually closely related to matrix factorization approaches as shown in [46]. Lastly, multiple random walk-based methods [58–60] combine the graph structure with additional node labels and attributes to achieve more informative embeddings and take advantage of fruitful nodal meta-data alongside the network structural information.

1.2.3 Deep learning based methods

Relatively recent pioneering works [61] have extended GRL to the deep learning theory, giving rise to Graph Neural Networks (GNN). Essentially, GNNs perform iterative message-passing extending convolution operations to graphs demonstrating remarkable performance by integrating node attributes and network structure during the embedding learning process. Convolution operations are defined over the local neighborhood of nodes while embedding aggregation is adopted to generalize the local structure and learn higher-level proximity. Graph Convolutional Neural Networks (GCNN) thereby scale linearly in the number of graph edges. Several examples of the representative power and success of GCNNs are given in [62, 63] and [64]. One of their limitations is usually the necessity for node features or else meta-data to avoid the over-smoothing pitfall hampering performance [42] when the GNN model defines deep architectures.

1.2.4 Latent space models

Latent Space Models (LSM) for the representation of graphs have been quite popular over the past years [65–71], especially for social networks analysis [72, 73] facilitating community extraction [74] and characterization of network polarization [75]. LSMs utilize the generalized linear model framework to obtain informative latent node embeddings while preserving network characteristics. The choice of latent effects in modeling the link probabilities between the nodes leads to different expressive capabilities characterizing network structure. In particular, in the Latent Distance Model (LDM) [76] nodes are placed closer in the latent space if they are similar or vice-versa. LDM obeys the triangle inequality and thus naturally represents transitivity [77, 78] (*"a friend of a friend is a friend"*) and network homophily [79, 80] (*a tendency where similar nodes are more likely to connect to each other than dissimilar ones*). Homophily is a very well-known and well-studied effect appearing in social networks [77,79,80] and essentially describes the tendency for people to form connections with those that share similarities with themselves. Similarities can be drawn from meta-data (observed node attributes) and may refer to shared demographic properties, political opinions, etc. Homophily has been observed among a broad range of collaborations (see [78] for a complete overview). Homophily can also be accounted for based on the unobserved attributes as defined by the LDM as shown in [81].

Homophily explains prominent patterns as expressed in social networks in terms of transitivity, as well as, balance theory ("the enemy of my friend is an enemy") [82].

1.3 Graph Representation Learning for temporal networks

So far, we have treated networks as static in time. Many networks evolve through time and are liable to modifications in structure with newly arriving nodes or emerging connections. GRL methods have primarily addressed static networks, in other words, a snapshot of the networks at a specific time. However, recent years have seen increasing efforts toward modeling dynamic complex networks, see also [73] for a review. Whereas most approaches have concentrated their attention on discrete-time temporal networks, which have built upon a collection of time-stamped networks [83–87] modeling of networks in continuous-time has also been studied [88–91]. These approaches have been based on latent class [88–90, 92, 93] and latent feature modeling approaches [66, 73, 91, 94–96], including advanced dynamic graph neural network representations [97, 98].

In this thesis, we will focus on a special class of dynamic networks characterized by a single event occurring between dyads, which we denote as single-event networks (SEN). I.e., links occur only once. Specifically, we aim to analyze citation networks which are a prominent example of SEN, as edges can appear only once at the time of the paper's publication. However, neither of the existing dynamic network modeling approaches explicitly account for SENs. Whereas continuous-time modeling approaches are designed for multiple events, thereby easily over-parameterize such highly sparse networks, static networks can easily be applied to such networks by disregarding the temporal structure but thereby potentially miss important structural information given by the event time. Despite these limitations, to the best of our knowledge, existing generative dynamic network modeling approaches do not explicitly account for single-event occurrences. In Figure 1.1, we provide an example of three cases of networks that define static, traditional event-based dynamic networks as well as single-event networks. We here observe how static networks are completely blind to the temporal information that single-event networks capture while it is also evident that they differ from traditional temporal networks where each dyad can have multiple events across time.

1.4 Graph Representation Learning and network science

Network Science and GRL are two disciplines that both operate in the realm of graph-structured data but focus on different aspects and utilize distinct method-

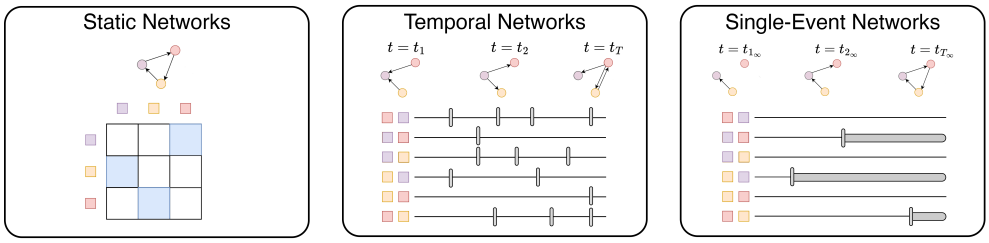


Figure 1.1. Examples of three different types of networks based on their temporal structure. Round points represent network nodes, square points make up the corresponding colored node dyads, arrows represent directed relationships between two nodes, vertical lines represent events, and black lines are the timelines while grey bold lines show that a link (event) appeared once and cannot be observed again. *Left panel:* Static networks where links occur once and there is no temporal information available. *Middle panel:* Temporal networks where links are events in time and can be observed multiple times along the timeline. *Right panel:* Single-event networks where links appear in a temporal manner but they can occur once, defining edges as single events.

ologies. Network Science focuses on the introduction of heuristics allowing for the analysis of network topology, clustering, centrality, community detection, and network dynamics. In contrast, GRL focuses on converting graph-structured data into (usually) continuous vector representations obtained via a learning procedure, and most importantly it supports predictive modeling. In summary, while both Network Science and GRL operate on graph-structured data, they differ in their objectives, methodologies, applications, data focus, and interdisciplinary roots. Network Science is more concerned with the analysis and understanding of complex networks, whereas GRL focuses on learning representations to facilitate predictive modeling and machine learning tasks. The main focus of this thesis will be the development of efficient GRL methods capable of the analysis of large-scale graphs.

1.5 Contributions

The central aim of this thesis is the development of novel algorithmic approaches for Graph Representation Learning by utilizing the Euclidean distance metric, under the Latent Distance Model formulation [76]. As a result, the proposed frameworks will obey the triangle inequality and naturally represent homophily and transitivity in the latent space, modeling high-order node proximity. It will be evident that such a choice leads to ultra-low dimensional graph representations, showcasing surprisingly superior performance when compared to multiple state-of-the-art models. Furthermore, the thesis will focus on structural-aware embeddings leading to hierarchical expressions, community characterization, and the discovery of extreme profiles in networks. Importantly the various presented methods will define unified learning pro-

cesses avoiding heuristics and multi-stage processes (e.g. post-processing steps). We will focus on a journey seeking unified network embeddings sufficient and powerful enough to characterize the structure, as well as, to successfully perform the multiple and different tasks graph analysis has to offer. This comes contrary to most state-of-the-art studies where models are designed to perform one or two tasks maximum without requiring two-level strategies and procedures for performing additional tasks. Our efforts will initially focus on positive integer weighted graphs, later we will extend the analysis to signed integer weighted graphs, as well as, single-event temporal networks.

The main contributions of the work can be summarized as follows:

- We introduce novel representation learning models over graphs for the study of both signed and unsigned, as well as, single-event networks.
- We learn node embeddings capable of extracting the hierarchical structure of the network at different scales, accounting for the community discovery, and extracting distinct profiles of networks.
- We, for the first time, define a likelihood function capable of the principled analysis of single-event temporal networks, while also quantifying nodal impact through modeling node receiving edge dynamics.
- We account for the computational costs of modern networks that can have millions or even billions of nodes. For that, we rely on accurate linearithmic approximations of the likelihood, unbiased random sampling procedures, and case-control inferences.
- We consider ultra low-dimensional node embeddings that are learned for moderate and large-scale networks and show high performance in all of the considered downstream tasks.
- We further highlight how the inferred hierarchical organization, community extraction, archetypal characterization, impact quantification, and low-dimensional representations can facilitate the visualization of network structures with high accuracy without requiring additional post-processing tools.
- The proposed frameworks are extended to the case of bipartite networks, where characterization of structure, hierarchical representations, community detection, and archetype extraction are considered arduous tasks, especially for signed networks.
- Extensive experimental evaluations demonstrate that the proposed approaches generally surpass widely adapted baseline methods in node classification, link prediction, and network reconstruction tasks.
- Importantly, the proposed frameworks define optimal embeddings that are characterized by the most consistent performance across different downstream tasks when compared to various prominent baselines.

- Lastly, this thesis aims to fill a missing part in the GRL literature which is to extensively position and benchmark the performance of Latent Distance Models for Graph Representation Learning against state-of-the-art baselines, showcasing their superior performance in multiple settings.

1.6 Organization

In Part I, we started with a brief introduction to Graph Representation Learning and network analysis as championed in the past years, followed by a methods chapter introducing the developed frameworks to be used throughout this thesis. We then continue with Part II of the thesis which addresses scalability, hierarchical representations, and community extraction in positive integer weighted networks. We showcase the performance of the proposed frameworks in downstream tasks utilizing ultra-low dimensions and importantly extend such analysis to positive integer-weighted bipartite networks. Afterward, Part III focuses on the analysis of signed integer weighted networks utilizing for the first time the Skellam distribution in network analysis. Specifically, we will propose frameworks able to characterize network polarization and discover extreme profiles and distinct aspects as present in networks. This will come as a generalization of Archetypal Analysis (AA) to relational data under both a defined latent space constrained to the convex hull of the latent embeddings, as well as, a Minimum Volume (MV) approach over the latent space. Part IV aims at the analysis of single-event temporal networks, combining an Inhomogeneous Poisson Point Process and the Latent Distance Model, defining a Single-Event Poisson Process. Lastly, Part V provides a discussion chapter based on the introduced theory, experiments, and results while the conclusion chapter concludes this thesis.

More detailed the thesis will follow a structure as presented below:

- **Part I: Introduction and methods**
  - Chapter 1: Introduction. This chapter provides a brief introduction to Graph Representation Learning and network analysis as championed in the past years.
  - Chapter 2: Methods. This chapter introduces the proposed methods for Graph Representation Learning to be used throughout this thesis.
- **Part II: Graph Representation Learning of positive integer weighted networks**
  - Chapter 3: *"A Hierarchical Block Distance Model for Ultra Low-Dimensional Graph Representations"*. This chapter is based on the original paper [99] currently accepted for publication by the journal *"IEEE Transactions on Knowledge and Data Engineering"*.

  - Chapter 4: *"HM-LDM: A Hybrid-Membership Latent Distance Model"*. This chapter is based on the original paper [74] published in the proceedings of *"The 11^{th} International Conference on Complex Networks and their Applications"*.

- **Part III: Graph Representation Learning of signed integer weighted networks**

  - Chapter 5: *"Characterizing Polarization in Social Networks using the Signed Relational Latent Distance Model"*. This chapter is based on the original paper [74] published in the proceedings of *"The 25^{th} International Conference on Artificial Intelligence and Statistics, AISTATS"*.

  - Chapter 6: *"A Hybrid Membership Latent Distance Model for Signed Integer Weighted Networks"*. This chapter is based on the original paper [100] published in the journal of *"Advances in Complex Systems"*.

- **Part IV: Graph Representation Learning of Single-event temporal networks**:

  - Chapter 6: *"Time to Cite: Modeling Citation Networks using the Dynamic Impact Single-Event Embedding Model"*. Preprint.

- **Part V Discussion and conclusion**:

  - Chapter 8: Discussion. This chapter discusses the general results, limitations, and future work of the thesis topic.
  - Chapter 9: Conclusion. This chapter concludes the thesis.

1.7 Reproducibility and code release

To enhance openness and reproducibility, the source code for all contributions presented in this thesis is publicly available and can be accessed in the following repositories:

- "A Hierarchical Block Distance Model for Ultra Low-Dimensional Graph Representations": *github.com/Nicknakis/HBDM*.

- "HM-LDM: A Hybrid-Membership Latent Distance Model": *github.com/Nicknakis/HM-LDM*.

- "Characterizing Polarization in Social Networks using the Signed Relational Latent Distance Model": *github.com/Nicknakis/SLIM_RAA*.
- "A Hybrid Membership Latent Distance Model for Unsigned and Signed Integer Weighted Networks": *github.com/Nicknakis/HM-LDM*.

CHAPTER 2

Methods

2.1 Notation

Before we start here we will briefly discuss the notation followed throughout the paper. We denote scalar values as lower-case and non-bold letters $\{x\}$, vectors are represented with lower-case bold letters $\{\boldsymbol{x}\}$, and matrices by upper-case bold letters $\{\boldsymbol{X}\}$. Single subscripts in lower-case bold letters $\{\boldsymbol{x}_i\}$ represent the $i'th$ vector while double subscripts in matrices $\{\boldsymbol{X}_{ij}\}$ denote the $i'th$ row and $j'th$ column single element of matrix $\{\boldsymbol{X}\}$.

2.2 What is a graph?

We will now provide a more formal definition of what we will consider a graph or a network, while we will use both of these terms interchangeably. Let now $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ define a graph with $\mathcal{V}$ being the vertex/node set and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ the edge set. We define an edge $(i, j) \in \mathcal{E}$ as the directed relationship having as source node $i \in \mathcal{V}$ and node $j \in \mathcal{V}$ as the target. In the following, we will focus on simple graphs with no loops and no multiple edges, meaning that there is no edge with the same source and target node and that every edge is uniquely defined. We will represent a graph by its adjacency matrix $\mathbf{Y}_{N\times N} = (y_{i,j})$ where $y_{i,j} = 0$ if the pair $(i, j) \notin \mathcal{E}$ otherwise it will be a non-zero value $y_{i,j} \neq 0$ for all $1 \leq i \leq N := |\mathcal{V}|$, and $1 \leq j \leq N := |\mathcal{V}|$. The most trivial case considers binary graphs where $y_{i,j} = 1$ if $(i, j) \in \mathcal{E}$, and $y_{i,j} = 0$ otherwise. We will characterize a graph as undirected when there are no directional relationships between the vertices, meaning that the edges do not have an inherent direction or arrow associated with them. In the undirected case, the adjacency matrix is symmetric, i.e. $\mathbf{Y} = \mathbf{Y}^\intercal$. In this thesis, we will initially focus on binary undirected graphs and later generalize to signed integer-weighted networks, assuming that the edge weights or the entries of the adjacency matrix can take any positive or negative integer value ($y_{ij} \in \mathbb{Z}$). In the case of the signed graphs, we will further denote $\mathcal{E}^+$ as the positive edge set meaning that $y_{i,j} > 0$ if the pair $(i, j) \in \mathcal{E}^+$, and accordingly $\mathcal{E}^-$ as the negative edge set with $y_{i,j} < 0$ if the pair $(i, j) \in \mathcal{E}^-$.

Lastly, as a special case, we will focus on bipartite graphs. A graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ will be a bipartite graph when the vertex set $\mathcal{V}$ can be partitioned into two non-empty and disjoint subsets $\mathcal{V}_1$ and $\mathcal{V}_2$ in such a way that every edge in $\mathcal{E}$ connects a vertex from

$\mathcal{V}_1$ to a vertex from $\mathcal{V}_2$. Formally now, for the vertex set $\mathcal{V}$, there exists a partition with $\mathcal{V} = \mathcal{V}_1 \cup \mathcal{V}_2$, and $\mathcal{V}_1 \cap \mathcal{V}_2 = \emptyset$ where $\mathcal{V}_1, \mathcal{V}_2$ being non-empty and disjoint sets. In addition, for any $(i, j) \in \mathcal{E}$, either $i \in \mathcal{V}_1$ and $j \in \mathcal{V}_2$, or $i \in \mathcal{V}_2$ and $j \in \mathcal{V}_1$.

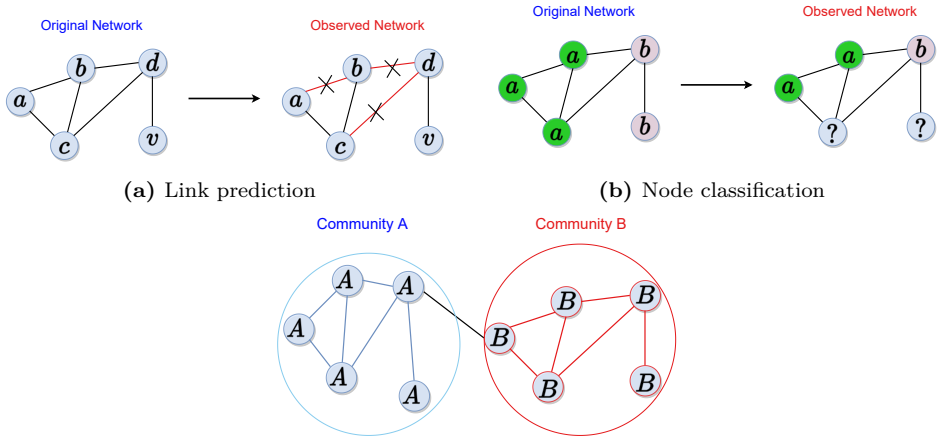
2.2.1 Downstream tasks

Machine learning on graphs focuses on characterizing emerging structures, discovering patterns, and extracting information that is present in graph-structured data [101]. The main goal of such an analysis/modeling of networks is the ability to perform and solve various important graph-related tasks. The most popular tasks, which will also be the main focus of this study, include relation/link prediction, node classification, and community detection.

Relation prediction or link prediction (we will use these terms interchangeably) refers to the task of predicting the existence or likelihood of certain relationships (edges) between node pairs in a graph. This task is particularly relevant in network analysis and has applications in various fields, including social networks [2], biological networks [102], and recommendation systems [103]. Formally, relation prediction can be defined as follows: given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ the goal is to predict whether a particular edge $(i, j) \in \mathcal{E}$ exists in the graph or to score the likelihood of the existence of different types of relationships between nodes i and j. The goal of relation prediction is to use the existing structure of the network and potentially additional features or attributes of nodes and edges to predict which links are most likely to form in the future or are currently missing from the network. This prediction task is usually formulated as either a binary classification problem, where the model predicts the presence or absence of a link between a pair of nodes, or as a ranking problem, where the model ranks the candidate links based on their likelihood score of existence. A toy example of such a task can be found in Figure 2.1 (a).

Node classification, also known as node attribute inference, is a fundamental problem in network analysis and machine learning on graphs [36, 38, 42]. It involves predicting the class or label of a node in a network based on its structural properties, attributes, and the labels of its neighboring nodes. The goal of node classification is to learn a predictive model that can generalize from labeled nodes (nodes with known classes) to classify unlabeled nodes existing in the network. This is typically done using supervised learning techniques, where the model is trained on a subset of nodes, i.e. $\mathcal{V}_{train} \subset \mathcal{V}$ with known labels and then used to predict the labels of a test or validation set of nodes in the network. Examples of node classification tasks include research topic prediction in citation networks [104], gene ontology type prediction in protein-protein interaction networks [105], and more [106]. A simple example of a node classification task can be found in Figure 2.1 (b).

Community detection, also referred to as node clustering, in the context of graph analysis, is the task of identifying groups of nodes in a network that are densely connected among themselves while having sparser connections to nodes in other groups [5, 107, 108]. These groups are often referred to as ”communities” or ”clus-



(a) Link prediction **(b)** Node classification

(c) Community detection

Figure 2.1. Downstream tasks for Graph Representation Learning. (a) Link prediction: In this setting, the network is partially observed and the task is to predict the missing links and regain the original network structure. (b) Node classification: In this setting, each network node has a label (in the example we have two labels a and b), the task is to infer the node labels for nodes with missing/unknown labels. (c) Community detection: In this setting, the whole network is observed and the task is to infer communities existing in the network (we show an example with two communities A and B).

ters" and their detection is essential for understanding the underlying structure and organization of complex networks. Community detection is a fundamental problem in network science and has numerous applications in various domains, such as social networks [109], opinion dynamics [110], protein-protein interactions [111], disease dynamics [112], and more [113]. By uncovering communities, researchers can gain insights into the modular organization and functional units within a network, which can lead to a better understanding of its behavior and facilitate targeted analyses and even interventions. Essentially, the task of community detection is to infer underlying latent structures or "communities" having only the network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ as an input. A community detection example can be seen in Figure 2.1 (c).

In terms of the "classical" machine learning theory, node classification and relation prediction are often categorized as semi-supervised tasks since they work both with labeled and unlabeled data during inference. Specifically, for the test nodes/node pairs there exists information in terms of the nodes' neighborhood in the graph which differs from the traditional supervised setting where test data are completely unobserved. Community detection and clustering are considered the unsupervised extension of classical clustering tasks to network data. One major difference when working with graph data is that the assumption of independent and identically distributed data (i.i.d.) does not hold since nodes in any network are interconnected and thus dependent. It is worth mentioning the existence of inductive biases when

modeling networks from different domains and disciplines, making the modeling of graphs more challenging than traditional machine learning problems.

So far, we have discussed node-level downstream tasks since they will be the main focus of this thesis. Nevertheless, there exist various important and interesting graph-level tasks. These include graph classification, regression, and clustering [34]. In these cases, the inputs to the learning procedure can potentially be a set of graphs while the predictions, rather than focusing on a single graph, are generalized to multiple different graphs.

2.3 Latent Space Models

Latent Space Models (LSMs) for the representation of graphs have been well established over the past years [65–71]. LSMs utilize the generalized linear model framework to obtain informative latent node embeddings while preserving network characteristics. The choice of latent effects in modeling the link probabilities between the nodes leads to different expressive capabilities for characterizing the network structure. Popular choices include the Latent Distance Model [76] which defines a probability of an edge based on the Euclidean distance of the latent embeddings, and the Latent Eigen-Model [114] which generalizes stochastic blockmodels, as well as, distance models. Various non-Euclidean geometries of LSMs have also been studied in [71] with the hyperbolic case being of particular interest [115].

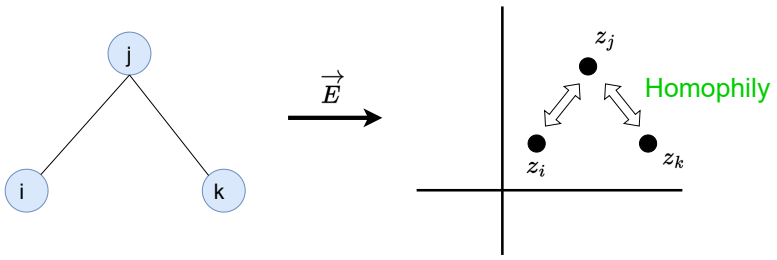


Figure 2.2. Expression of homophily and transitivity as imposed by the Latent Distance model. Black lines correspond to network edges. Connected nodes are positioned close to each other to define a high probability of an edge, e.g. pairs $\{i, j\}$ and $\{j, k\}$. Consequently, the distance of node pair $\{i, k\}$ is bounded by the triangle inequality, and thus node pair $\{i, k\}$ has to also be positioned in close proximity.

2.3.1 Latent Distance Models

Here, we will focus on the Latent Distance Model (LDM) [76] where network nodes are positioned close in the latent space if they are connected or share additional similarities, such as high-order dependency or proximity. Most importantly, the LDM obeys the triangle inequality and thus naturally represents transitivity and homophily. The LDMextends the traditional homophily expression to the case where no node meta-data exist through the introduction of unobserved attributes as defined by the LDM [81]. Specifically, we define a D-dimensional latent space ($\mathbb{R}^D$) in which every node of the graph is characterized through the unobserved but informative node-specific variables $\{\mathbf{z}_i \in \mathbb{R}^D\}$. These variables are considered sufficient to describe and explain the underlying relationships between the nodes of the network. The probability of an edge occurring is considered conditionally independent given the unobserved latent positions and depends on the Euclidean distance. Consequently, the total probability distribution of the network can be written as:

$$P(Y|\boldsymbol{Z}, \boldsymbol{\theta}) = \prod_{i<j}^{N} p(y_{i,j}|\mathbf{z}_i, \mathbf{z}_j, \boldsymbol{\theta}_{i,j}), \tag{2.1}$$

where $\boldsymbol{\theta}$ denotes any potential additional parameters, such as covariate regressors. A popular and convenient parameterization of Equation (2.1) for binary data is through the logistic regression model [76, 81, 116, 117].

For our study, we will focus on an LDM under the Poisson distribution [117]. The Poisson LDM generalizes the analysis to integer-weighted graphs while the exchange of the *logit* to an *exponential* link function when transitioning from a Bernoulli to a Poisson model defines nice decoupling properties over the predictor variables in the likelihood [118, 119]. Importantly, we will also study binary networks where the use of a Poisson likelihood for modeling such relationships in a network does not decrease the predictive performance nor the ability of the model to detect the network structure [120]. Formally now, we define the Poisson rate of the LDM for an occurring edge based on the Euclidean distance between the latent positions of the two nodes as:

$$\lambda_{ij} = \exp\big(\gamma_i + \gamma_j - d(\mathbf{z}_i, \mathbf{z}_j)\big). \tag{2.2}$$

In this formulation, we consider the LDM Poisson rate with node-specific biases or random effects [81, 117]. In particular, $\gamma_i \in \mathbb{R}$ denotes the node-specific random effect and $d_{ij}(\mathbf{z}_i, \mathbf{z}_j) = ||\mathbf{z}_i - \mathbf{z}_j||_2$ denotes the Euclidean distance (or potentially any distance metric obeying the triangle inequality) $\big\{d_{ij} \leq d_{ik} + d_{kj},\ \forall (i,j,k) \in V^3\big\}$. Considering variables $\{\mathbf{z}_i\}_{i \in V}$ as the latent characteristics, Equation (2.2) shows that similar nodes will be placed closer in the latent space, yielding a high probability of an occurring edge and thus modeling homophily and satisfying network transitivity and reciprocity through the triangle inequality. Essentially, we extend the meaning of similarity to the unobserved (latent) covariates, i.e., latent embeddings matrix $\mathbf{Z}$.

Connected or similar nodes define strong relationships that are to be translated by the LDM into the latent space, defining a high probability of observing connections. As a result, for two similar nodes $\{i, j\}$ the pairwise distance $||\mathbf{z}_i - \mathbf{z}_j||_2$ should be small which further implies that for a different node $\{k\}$ we obtain $||\mathbf{z}_i - \mathbf{z}_k||_2 \approx ||\mathbf{z}_j - \mathbf{z}_k||_2$. The latter concludes that nodes $\{i, j\}$ are similar since they share similar relationships with the rest of the nodes. For an illustration please visit Figure 2.2. An immediate result of obeying the triangular inequality is that the LDM successfully models high-order interactions, as present in complex systems [121, 122]. The node-specific bias can account for degree heterogeneity, whereas the conventional LDM rate utilizing a global bias, γ^g, corresponds to the special case in which $\gamma_i = \gamma_j = 0.5\gamma^g$.

2.4 The Hierarchical Block Distance Model

The classical LDM, as introduced previously, naturally conveys the main motivation of Graph Representation Learning where similar nodes are positioned in close proximity in a constructed latent space. This comes as a direct consequence of the Euclidean metric choice, representing homophily, transitivity, and high-order proximity. Unfortunately, two equally important properties, scalability (analysis of large-scale networks is infeasible) and structure characterization are not met by the LDM. Specifically, it scales quadratically in terms of the number of network nodes as $\mathcal{O}(N^2)$ while it is agnostic in terms of latent structures that potentially exist in different scales.

Our goal is to design a Hierarchical Block Model preserving homophily and transitivity properties with a total complexity allowing for the analysis of large-scale networks. Similar to a classical Poisson LDM, we define the rate of a link between each network dyad $(i, j) \in V \times V$ based on the Euclidean distance, as shown in Equation (2.2). Such a decision guarantees that our model will inherit the natural properties of the so-effective LDM, and satisfy homophily and transitivity.

Moving to the next two properties, we can define a block-alike hierarchical structure by a divisive clustering procedure over the latent variables in the Euclidean space. Incorporating a block structure into the model facilitates the retrieval of underlying structures, while the integration of a hierarchy accounts for the emergence of these structures across multiple scales. In addition, we will constrain the total optimization cost of such a model to a linearithmic upper bound complexity, making large-scale analysis feasible. We will start such a procedure by initially noticing that a shallow clustering of the latent space with a number of clusters, K, equal to the number of nodes, N, leads to the same log-likelihood as of the standard LDM, defining a sum over each ordered pair of the network, as:

$$\log P(\mathbf{Y}|\mathbf{\Lambda}) = \sum_{\substack{i<j \\ y_{ij}=1}} \log(\lambda_{ij}) \; - \; \sum_{i<j} \Big(\lambda_{ij} + \log(y_{ij}!)\Big) \, . \tag{2.3}$$

where $\mathbf{\Lambda} = (\lambda_{ij})$ is the Poisson rate matrix which has absorbed the dependency over the model parameters while we presently ignore the linear scaling by dimensionality

D of the above log-likelihood function. Notably, the first term of Equation (2.3), which hereby we will refer to as link contribution/term $\sum_{y_{i,j}=1} \log(\lambda_{i,j})$, is responsible for positioning "similar" nodes closer in the latent space, expressing the desired homophily. This is straightforward by substituting Equation (2.2) in the link contribution that is maximized when the distance is zero (for fixed random effects and fixed latent embeddings for all nodes except nodes $\{i, j\}$). The second term of Equation (2.3) $\sum_{i<j} \lambda_{ij}$, from now on referred to as the non-link contribution/term, acts as the repelling force for dissimilar nodes, being responsible for positioning nodes far apart, and in the case of $y_{ij} = 0$ is maximized when $d_{ij} \to +\infty$ (by fixing again the rest of parameters).

Focusing on the computational complexity of Equation (2.3), and given that large networks are highly sparse [15] with the number of edges proportional to the number of nodes in the network, results in a low computation cost the link contribution. We empirically showed in Figure 2.3 that it scales linearithmic or sub-linearithmic with N. Importantly, the link term removes rotational ambiguity between the different blocks/clusters of the hierarchy (as discussed later). For these reasons, no block structure is imposed on the calculation of the link contribution.

Moving now to the non-link term, it requires the computation of all node pairs distance matrix and thus it scales as $\mathcal{O}(N^2)$ making the evaluation of the above likelihood infeasible for large networks and being the main overhead for both space and time complexities. In order to make such a calculation linearithmic, we aim to enforce a block structure, i.e., akin to stochastic block models [124–126], when grouping the nodes into K clusters we define the rate between block k and k' in terms of their distance between centroids. We initialize such a procedure, by a shallow block structure obtaining the following non-link expression:

$$\sum_{i<j} \lambda_{ij} \approx \sum_{k=1}^{K} \Bigg(\sum_{\substack{i<j \\ i,j \in C_k}} \exp\big\{\big(\gamma_i + \gamma_j - ||\mathbf{z}_i - \mathbf{z}_j||_2\big)\big\} + \sum_{k'>k}^{K} \sum_{i \in C_k} \sum_{j \in C_{k'}} \exp\big\{\big(\gamma_i + \gamma_j - ||\boldsymbol{\mu}_k - \boldsymbol{\mu}_{k'}||_2\big)\big\} \Bigg), \tag{2.4}$$

where $\boldsymbol{\mu}_k$, has absorbed the dependency over the variables $\boldsymbol{Z} \in \mathbb{R}^{N \times D}$, and denotes the k'th cluster centroid over the set of K total centroids $\boldsymbol{C} = \{C_1, \ldots, C_K\}$. Cluster centroids $\boldsymbol{\mu}_k$ are implicit parameters defined as a function over the latent variables, as it will be clear later. In general, the clustering procedure is expected to naturally extend the concept of homophily to the level of clusters via the centroid expressions. This means that on a node level, closely related nodes will be grouped together in clusters while on a cluster level interconnected clusters will also be positioned closely in the latent space, creating an effective block structure representation. Overall, the clustering technique adheres to "cluster-homophily" and "cluster-transitivity" within the latent space.

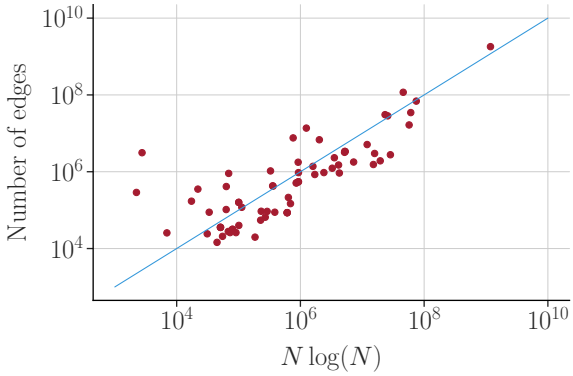
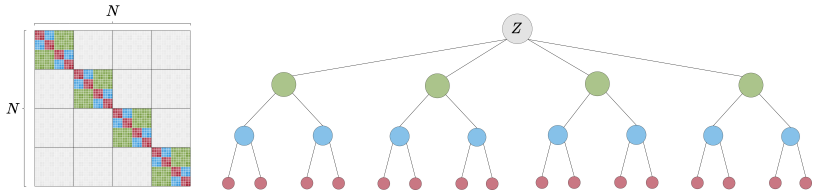
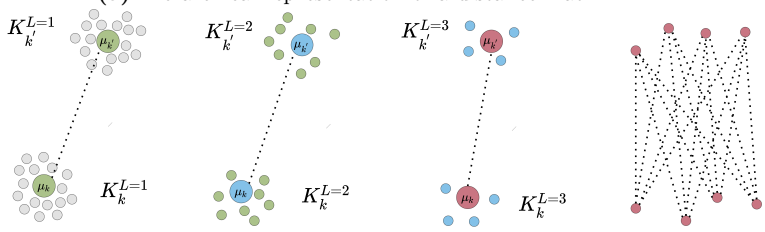


Figure 2.3. Log-Log plot of the number of network edges versus $N \log N$ where N the number of vertices, for 70 datasets of the SNAP library [123].



(a) Hierarchical representation of a distance matrix



(b) Pairwise distance approximation

Figure 2.4. Schematic representation of the distance matrix calculation for a hierarchical structure of the tree of height $L = 3$ and for the number of observations $N = 64$. (a) Hierarchical representation of the all-pairs distance matrix. (b) Pairwise distance approximation based on cluster centroids across different levels of the hierarchy [99].

2.4.1 A Hierarchical Representation

To attain the required hierarchical organization, we utilize hierarchical clustering via a divisive procedure. In more detail, we organize the embedded clusters into a hierarchical tree structure, forming a cluster dendrogram. The tree's root defines a single cluster containing all latent variable embeddings $\boldsymbol{Z}$. During the construction of the tree, clusters (tree-nodes) are divided at each level until every tree-node becomes

a leaf. A cluster is considered a leaf node if it contains an equal or smaller number of network nodes than an established threshold, N_{leaf}. The threshold value is chosen in such a way that to maintain linearithmic efficiency in terms of complexity and is set to $N_{\text{leaf}} = \log N$, which leads to roughly $K = N/\log(N)$ total clusters. The tree-nodes belonging to a specific tree-level are considered the clusters for that specific tree height. Every new division of a non-leaf node is performed solely on the set of points assigned to the parent node in the tree (tree-node/cluster). At each level of the tree, the distance between corresponding cluster centroids is considered as the pairwise distances of datapoints that belong to different tree-nodes, as shown in Figure 2.4 (ii). Utilizing these distances, we compute the likelihood contribution defined by the blocks and proceed with binary divisions, moving down the tree, for the nodes that are not leaf nodes. When every tree-node is regarded as a leaf, we analytically determine the inner cluster pairwise distances for the corresponding likelihood contribution of the analytical blocks, as depicted in the final part of Figure 2.4 (ii). This analytical calculation is carried out at a linearithmic cost of $\mathcal{O}(KN_{\text{leaf}}^2) = \mathcal{O}(N \log N)$, and it reinforces the homophily and transitivity characteristics of the model. Specifically, for network nodes that are most similar, the model calculates explicitly the latent distances in the same manner as the standard LDM.

We can thereby define a Hierarchical Block Distance Model with Random Effects (HBDM-Re) as:

$$
\begin{aligned}
\log P(Y|\mathbf{Z}, \boldsymbol{\gamma}) = & \sum_{\substack{i<j \\ y_{i,j}=1}} \Big(\gamma_i + \gamma_j - ||\mathbf{z}_i - \mathbf{z}_j||_2 \Big) \\
& - \sum_{k=1}^{K_L} \Bigg(\sum_{\substack{i<j \\ i,j \in C_k^{(L)}}} \exp\{(\gamma_i + \gamma_j - ||\mathbf{z}_i - \mathbf{z}_j||_2)\} \Bigg) \\
& - \sum_{l=1}^{L} \sum_{k=1}^{K_l} \sum_{k'>k}^{K_l} \Bigg(\exp\Big\{(-||\boldsymbol{\mu}_k^{(l)} - \boldsymbol{\mu}_{k'}^{(l)}||_2)\Big\} \\
& \times \Big(\sum_{i \in C_k^{(l)}} \exp\{\gamma_i\} \Big) \Big(\sum_{j \in C_{k'}^{(l)}} \exp\{\gamma_j\} \Big) \Bigg),
\end{aligned} \tag{2.5}
$$

where $l \in \{1, \ldots, L\}$ denotes the l'th dendrogram level, k_l is the index representing the cluster id for the different tree levels, and $\boldsymbol{\mu}_k^{(l)}$ the corresponding centroid. We also consider a Hierarchical Block Distance Model (HBDM) without the random effects which is achieved by setting $\gamma_i = 0.5\gamma^g$. For a multifurcating tree that splits into K clusters and has $N/\log(N)$ terminal nodes or clusters, there are $\mathcal{O}(N/(K \log N))$ internal nodes. Each node requires the evaluation of $\mathcal{O}(K^2)$ pairs, leading to an overall complexity of $\mathcal{O}(NK/\log N)$. Therefore, K must be less than or equal to $\log N^2$ to achieve a scaling of $\mathcal{O}(N \log N)$ [127]. It's noteworthy to observe that in

Equation (2.5), the random effects contributing to the non-link term are independent of the centroid distance calculations. As a result, the selection of the exponential link function allows an implicit calculation over the pairwise rates of the approximation term, facilitating efficient computations.

2.4.2 Divisive partitioning using k-means with a Euclidean distance metric

The likelihood formula provided by Equation (2.5) can be minimized directly by allocating nodes to clusters given the tree structure. Unfortunately, performing such an evaluation for all N nodes results in a scaling that becomes impractical, defining a $\mathcal{O}(N^2/\log N)$ complexity. To make this more manageable, we employ a more efficient method of divisive partitioning, which minimizes the Euclidean norm $||\boldsymbol{\mu}_{k_l} - \boldsymbol{\mu}_{k'_l}||_2$. The divisive clustering procedure thus relies on the following Euclidean norm objective

$$J(\mathbf{r}, \boldsymbol{\mu}) = \sum_{i=1}^{N} \sum_{k=1}^{K} r_{ik} ||\mathbf{z}_i - \boldsymbol{\mu}_k||_2, \tag{2.6}$$

where k denotes the cluster id, $\mathbf{z}_i$ is the i'th data observation, r_{ik} the cluster responsibility/assignment, and $\boldsymbol{\mu}_k$ the cluster centroid.

The given objective function is not supported by existing k-means clustering algorithms that depend only on the squared Euclidean norm. As a result, we now develop an optimization procedure specifically for k-means clustering under the Euclidean norm. This approach lies within the auxiliary function framework as developed in the context of compressed sensing in [128]. We establish an auxiliary function for (2.6) as follows:

$$J^{+}(\boldsymbol{\phi}, \mathbf{r}, \boldsymbol{\mu}) = \sum_{i=1}^{N} \sum_{k=1}^{K} r_{ik} \left(\frac{||\mathbf{z}_i - \boldsymbol{\mu}_k||_2^2}{2\phi_{ik}} + \frac{1}{2}\phi_{ik} \right), \tag{2.7}$$

where $\boldsymbol{\phi}$ are the auxiliary variables. Thereby, minimizing Equation (2.7) with respect to $\boldsymbol{\phi}_{nk}$ yields $\boldsymbol{\phi}^*_{ik} = ||\mathbf{z}_i - \boldsymbol{\mu}_k||_2$ and by plugging $\boldsymbol{\phi}^*_{ik}$ back to (2.6) we obtain $J^{+}(\boldsymbol{\phi}^*, \mathbf{r}, \boldsymbol{\mu}) = J(\mathbf{r}, \boldsymbol{\mu})$ verifying that (2.7) is indeed a valid auxiliary function for (2.6). The algorithm proceeds by optimizing cluster centroids as

$$\boldsymbol{\mu}_k = \Big(\sum_{i \in k} \frac{\mathbf{z_i}}{\phi_{ik}} / \sum_{i \in k} \frac{1}{\phi_{ik}} \Big), \tag{2.8}$$

and assigning points to centroids as

$$\arg\min_{\boldsymbol{C}} = \sum_{k=1}^{K} \sum_{\boldsymbol{z} \in C_k} \Big(\frac{||\boldsymbol{z} - \boldsymbol{\mu}_k||_2^2}{2\boldsymbol{\phi}_k} + \frac{1}{2}\boldsymbol{\phi}_k \Big), \tag{2.9}$$

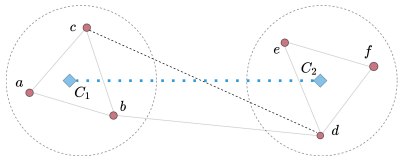
upon which $\boldsymbol{\phi}_k$ is updated. The overall complexity of this procedure is $\mathcal{O}(TKND)$ [129] where T is the number of iterations required to converge. As shown in [128], Equation (2.7) is a special case of a general algorithm for an $l_p(0 < p < 2)$ norm minimization using an auxiliary function with the algorithm converging faster the smaller p is. For a detailed study of the efficiency of the optimization procedure under such an auxiliary function, see [128].

Number of splits in each layer of the divisive procedure: A straightforward approach to constructing the tree structure would be via an agglomerative procedure where essentially the nodes would be split into $K = N/\log(N)$ followed by binary merges until only one cluster survives. Despite this being possible under the above Euclidean k-means procedure, it scales prohibitive and thus does not respect the linearithmic complexity threshold. For that, we turn to a divisive clustering procedure for constructing the dendrogram. In such a procedure, lies a trade-off between the number of nodes belonging to each cluster and the distance approximation quality. It is evident that an initial binary split would be a very crude distance approximation and as a result we choose in the initial split to create the maximum allowed number of clusters respecting the linearithmic complexity threshold $\mathcal{O}(N \log N)$, that is equal to $K = \log N$. Continuing to divide into $\log N$ clusters might seem like an appealing approach, but for a balanced multifurcating tree that has $N/\log N$ leaf clusters, this strategy would lead to a height scaling of $\mathcal{O}(\log N/\log\log N)$. Consequently, the overall complexity of this method would be $\mathcal{O}(N \log^2(N)/\log\log N)$ [127]. A balanced binary tree at all levels beneath the root leads to a height scaling of $\mathcal{O}(\log N)$, with each level of the tree accounting for $\mathcal{O}(DN)$ operations. When including the linear scaling factor due to dimensionality D, this results in an overall complexity of $\mathcal{O}(DN \log N)$. Figure 2.4 (i) depicts the resulting tree for a small problem involving $N = 64$ nodes. In this example, the nodes are first divided into 4 clusters (approximately equal to $\log(64)$), and then binary splits are performed until each leaf cluster contains 4 nodes (also roughly equivalent to $\log(64)$)[1].

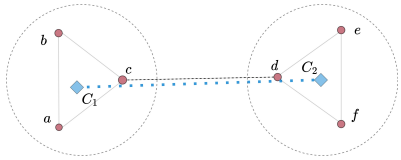
2.4.3 Hierarchical Block Representations Expressing Homophily and Transitivity

A crucial aspect in maintaining the homophily and transitivity properties of HBDM-Re and HBDM is to avoid approximating the link terms at the block level, as is done in (hierarchical) SBMs. Instead, the link contribution to the log-likelihood across the entire hierarchy should be calculated analytically, going beyond the leaf/analytical blocks. Figure 2.5 (a) and (b) depict two leaf clusters connected by a link. Suppose that the distances within the blocks are computed analytically and that both the link and non-link contributions of pairs across different clusters are estimated based on the distance between their centroids. Such an approach would essentially permit any rotation of each cluster throughout the hierarchy, as neither the inner-block

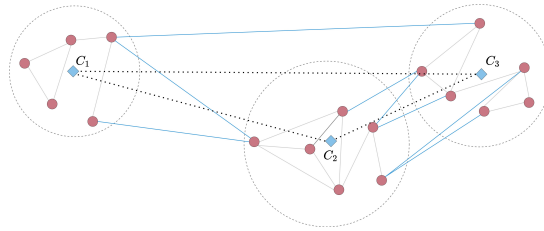
[1]For visualization purposes only, we show equally sized clusters.



(a) Non-optimal rotation over leaf clusters.



(b) Optimal rotation over leaf clusters.



(c) Three latent block structures.

Figure 2.5. The clusters within the dashed circles denote the leaf block structures. The red circles and blue rhombuses indicate the node embeddings and the centroids, respectively. Gray lines represent the links and the dashed lines the distance between the cluster centers [99].

distances (analytical) nor the centroid distances would be altered by these rotations, resulting in an identical likelihood, showcasing a block-level rotational invariance. In this scenario, homophily expression would be compromised. For instance, the distance between nodes c and d might not necessarily be shorter than the distance between other disconnected inter-cluster pairs (e.g., as shown in Figure 2.5 (a)). This illustrates how the rotation of the blocks can have a significant effect on the homophily characteristics of HBDM-Re and HBDM.

Computing the link contributions between different clusters analytically resolves this ambiguity, as the likelihood is penalized more when nodes c and d are positioned in a way that is not aware of rotations. The computational cost of taking into account all the link terms analytically means that the model's complexity is tied to the number of network edges contrary to a total block structure where the complexity is strictly linearithmic. However, we demonstrate empirically in Figure 2.3 that the number of network edges E scales linearly with $N \log N$, so this analytical term complies with our complexity boundary. Figure 2.5 (c) illustrates examples of clusters that define cases of block interconnections between both sparsely connected blocks ($\{C_1, C_3\}$, $\{C_1, C_2\}$) and densely connected blocks ($\{C_2, C_3\}$). The analytical links between clusters (depicted as blue lines) are instrumental in determining the proper orientation of the blocks. Furthermore, these inter-cluster links guide the proximities of the centroids at the cluster level, thus playing a vital role in upholding the properties of cluster homophily and transitivity.

In the HBDM, pairwise distances remain unaffected by rotation, reflection, and translation operations of the latent space due to its inheritance from the LDM [76] (even though the dyad rates λ_{ij} are uniquely defined). These isometries can be

addressed through a Singular-Value-Decomposition procedure. Analytically, let $\boldsymbol{Z}$ denote the embedding of our proposed HBDM(-RE) such that the i'th row $(\boldsymbol{Z})_i = \mathbf{z}_i$. Then, visualizations of the inferred latent space can be uniquely determined by imposing a centering step $\widehat{\boldsymbol{Z}} = \boldsymbol{Z} - \bar{\boldsymbol{Z}}$ followed by a singular value decomposition of the latent positions $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T = \text{SVD}(\widehat{\boldsymbol{Z}})$ to remove rotation ambiguity. Thereby, we can introduce $(\boldsymbol{Z}^*)_i = (\boldsymbol{U}\boldsymbol{\Sigma})_i$ which determines uniquely identifiable latent positions as long as the singular values are distinct.

While the analytical calculation of link terms introduces rotational awareness to the HBDM clusters, we further explore the conditions under which a continuous operation that defines infinitesimal rotations (relative to the cluster centroid) is permissible. This exploration seeks to understand the situations in which the loss function of Equation (2.5) remains invariant to continuous rotations. In Lemma 2.4.1 (proof follows), we start our investigation of this problem by showing that blocks with a unique inter-cluster link connection reduce the clusters' degree of rotational freedom by one.

Lemma 2.4.1. *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a graph and let $\mathcal{C}$ be a cluster with its centroid located at $\boldsymbol{\mu} \in \mathbb{R}^D$ having an edge $(i, j) \in \mathcal{E}$ for some $i \in \mathcal{C}$ and $j \in \mathcal{V} \backslash \mathcal{C}$ such that $\mathbf{z}_i \neq \boldsymbol{\mu}$. If $\tilde{\mathbf{z}}_i = \boldsymbol{\mu} + \mathbf{R}(\boldsymbol{\theta})(\mathbf{z}_i - \boldsymbol{\mu})$ such that $\mathbf{R}(\boldsymbol{\theta})$ is a rotation matrix acting on the embeddings of nodes in cluster $\mathcal{C}$, then the maximum degree of freedom of any infinitesimal λ_{ij}-invariant rotation is defined by $\boldsymbol{\theta} \in \mathbb{R}^{D-2}$.*

Proof. A general rotation matrix, $\mathbf{R}(\boldsymbol{\theta})$, for a D-dimensional space is given by a rotation angle vector $\boldsymbol{\theta} \in \mathbb{R}^{D-1}$. Define $\tilde{\lambda}(\boldsymbol{\theta})_{ij} = \exp\left(\gamma_i + \gamma_j - \|\tilde{\mathbf{z}}_i - \mathbf{z}_j\|\right)$ such that $\tilde{\lambda}(\mathbf{0})_{ij} = \lambda_{ij}$. Then, an infinitesimal rotation leaving λ_{ij} unchanged must be along the direction of a non-zero vector $\boldsymbol{v} \in \mathbb{R}^{D-1}$ requiring $\langle \frac{\partial \tilde{\lambda}(\boldsymbol{\theta})_{ij}}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\mathbf{0}}, \boldsymbol{v} \rangle = 0$. This equation is satisfied either if (i) $\frac{\partial \tilde{\lambda}(\boldsymbol{\theta})_{ij}}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\mathbf{0}} = \mathbf{0}$, which would require either $||\mathbf{z}_i - \mathbf{z}_j||$ is maximum or minimum on the sphere defined by the rotation such that any infinitesimal rotation would respectively decrease or increase $\tilde{\lambda}_{ij}$; (ii) $\boldsymbol{v}$ is orthogonal to the gradient $\frac{\partial \tilde{\lambda}(\boldsymbol{\theta})_{ij}}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\mathbf{0}}$; consequently, this removes a degree of rotational freedom such that $\boldsymbol{\theta} \in \mathbb{R}^{D-2}$. □

An immediate implication of Lemma 2.4.1 is that in a two-dimensional embedding, continuous rotation of a cluster with only one external edge is not possible. For connected graphs, there is always a path from one node to all others, and thus every cluster must possess at least one external link. When considering the more general scenario of blocks with multiple inter-cluster edges, rotations that preserve the aggregate sum of pairwise distances among node embeddings become highly improbable, as elaborated in the next paragraph. As a result, for connected networks, we can generally anticipate the uniqueness of (local) minimum solutions, with no continuous rotations allowed that would leave the HBDM loss function of Equation (2.5) invariant.

As previously mentioned, local operations defined on the clusters with respect to their centroids can potentially leave the loss function value invariant since HBDM(-RE) calculates the non-link contributions between clusters in the objective function based on their centroids distances. It can be said that there are almost surely no infinitesimal local cluster rotations of local minima solutions of Equation (2.5) for 2-dimensional embeddings. We consider infinitesimal rotations on the cluster embeddings since our main motivation relies on the uniqueness of embeddings around the local minima so we also discard the local reflections of the clusters since they do not provide continuous transformation operations. Specifically, let $\mathcal{C}$ be a cluster with multiple external edges and let $\tilde{\mathbf{z}}_i = \boldsymbol{\mu} + \mathbf{R}(\boldsymbol{\theta})(\mathbf{z}_i - \boldsymbol{\mu})$ such that $\mathbf{R}(\boldsymbol{\theta})$ is a rotation matrix acting on the embeddings of nodes in cluster $\mathcal{C}$. We first note that any rotation of the cluster $\mathcal{C}$ by $\mathbf{R}(\boldsymbol{\theta})$ will leave all terms invariant in Equation (2.5) except for the sum over external edges $S = \sum_{(i,j)\in E_{\mathcal{C},V\setminus\mathcal{C}}} \log \lambda_{ij}$. For a local minima, no infinitesimal rotation exists that will reduce the overall sum of distances between node pairs defining the external edges as such rotation would improve the solution violating that it is a (local) minima. We can therefore assume that any rotation will result in either no change of the sum of distances or that the overall sum will increase. For embeddings in two-dimensional space, the rotations can be parameterized by the single parameter $\boldsymbol{\theta}$ that for a local minimum has the property $\frac{\partial S}{\partial \boldsymbol{\theta}} = \sum_{(i,j)\in E_r} \frac{\partial \log \tilde{\lambda}_{ij}}{\partial \boldsymbol{\theta}} - \sum_{(i,j)\in E_i} \frac{\partial \log \tilde{\lambda}_{ij}}{\partial \boldsymbol{\theta}} = 0$. As a result, edges reducing their distances (E_r) will have a positive gradient of $\boldsymbol{\theta}$ whereas edges increasing (E_i) will have a negative gradient, and these parts perfectly cancel out for a local minimum. However, as the overall sum of distances for the local minima cannot decrease, the overall sum must remain the same. Therefore, for all node pairs for which $(\mathbf{z}_i - \boldsymbol{\mu})^\top(\mathbf{z}_j - \boldsymbol{\mu}) > 0$, we have that the distance increases for increasing edges more than the reduction for decreasing edges. Furthermore for node pairs for which $(\mathbf{z}_i - \boldsymbol{\mu})^\top(\mathbf{z}_j - \boldsymbol{\mu}) < 0$, we, in general, expect further distances between edge pairs, thus less impact on the rotation. As a result, it is highly unlikely that clusters with more than one external edge can be rotated in two-dimensional space. As the likelihood in Equation (2.5) is defined on a connected network every cluster will have at least one external edge. In combination with Lemma 2.4.1, a local minima can therefore not be infinitesimally rotated.

2.4.4 A Hierarchical Block Distance Model for Bipartite Networks

Our proposed frameworks, HBDM and HBDM-RE have straightforward generalizations to both directed and bipartite graphs. In the following, we provide the mathematical extension for the bipartite case (the directed network formulation of our proposed model can be considered a special case of the bipartite framework in which self-links are removed and thus omitted from the below log-likelihood).

For a bipartite network with adjacency matrix $Y^{N_1\times N_2}$ we can formulate the log-likelihood as:

$$\begin{aligned}\log P(Y|\mathbf{\Lambda}) = & \sum_{\substack{i,j \\ y_{i,j}=1}} \Bigg(\psi_i + \omega_j - ||\mathbf{w}_i - \mathbf{v}_j||_2 \Bigg) \\ & - \sum_{k_L=1}^{K_L} \Bigg(\sum_{i,j \in C_{k_L}} \exp\{(\psi_i + \omega_j - ||\mathbf{w}_i - \mathbf{v}_j||_2)\} \Bigg) \\ & - \sum_{l=1}^{L} \sum_{k=1}^{K_l} \sum_{k'>k}^{K_l} \Bigg(\exp\Big\{(-||\boldsymbol{\mu}_k^{(l)} - \boldsymbol{\mu}_{k'}^{(l)}||_2)\Big\} \\ & \times \Big(\sum_{i \in C_k^{(l)}} \exp\{\psi_i\} \Big) \Big(\sum_{j \in C_{k'}^{(l)}} \exp\{\omega_j\} \Big) \Bigg), \end{aligned} \tag{2.10}$$

where $\{\boldsymbol{\mu}_k^{(l)}\}_{k=1}^{K_L}$ are the latent centroids that have absorbed the dependency of both sets of latent variables $\{\mathbf{w}_i, \mathbf{v}_j\}$ while we define the Poisson rate as:

$$\lambda_{ij} = \exp\big(\psi_i + \omega_j - d(\mathbf{w}_i, \mathbf{v}_j)\big), \tag{2.11}$$

where ψ_i and ω_j are the corresponding random effects and $\{\mathbf{w}_i,\ \mathbf{v}_j\}$ are the latent variables of the two disjoint sets of the vertex set of sizes N_1 and N_2, respectively. In this setting, we use our divisive Euclidean distance hierarchical clustering procedure over the concatenation $\mathbf{Z} = [\mathbf{W}; \mathbf{V}]$ of the two sets of latent variables. Therefore, we define an accurate hierarchical block structure for bipartite networks, with each block including nodes from both of the two disjoint modes. Here, a centroid is considered a leaf if the corresponding tree-cluster contains less than $\log(N_1)$ of the latent variables $\{\mathbf{w}_i\}_{i=1}^{N_1}$ or less than $\log(N_2)$ of $\{\mathbf{v}\}_{j=1}^{N_2}$.

2.4.5 Complexity Comparison

TABLE 2.1 offers a comparison of the time complexities for various notable GRL methods, expressed in Big $\mathcal{O}$ notation, akin to [130]. From this comparison, it becomes evident that our proposed HBDM model ranks among the most competitive frameworks. Regarding space complexity, our model exhibits linearithmic complexity, setting it apart from the majority of the considered baseline methods, which typically display quadratic space complexity, as shown in [130].

2.5 Hybrid memberships, Matrix Factorization, and Latent Distance Models

Revisiting our main goal which is to learn a representation in a lower dimensional space, expressing the property that similar nodes in the network are positioned closer

Table 2.1. Complexity analysis of methods. $N := |V|$ is the vertex set, $|E|$: edge set, $\mathcal{W}$: number of walks, $\mathcal{L}$: walk length, H: height of the hierarchical tree, D: node representation size, k: number of negative instances, q: order value, c: Chebyshev expansion order, γ: window size, α_1 and α_2 constants such as $\alpha_1, \alpha_2 \ll N$.

Method	Complexity
DeepWalk [22]	$\mathcal{O}\left(\gamma N \log\left(N\right)\mathcal{W}\mathcal{L}\mathcal{D}\right)$
Node2Vec [4]	$\mathcal{O}\left(\gamma N \mathcal{W}\mathcal{L}\mathcal{D}k\right)$
LINE [37]	$\mathcal{O}\left(\lvert E\rvert Dk\right)$
NetMF [46]	$\mathcal{O}\left(N^2 D\right)$
NetSMF [47]	$\mathcal{O}\left(\lvert E\rvert(\gamma + D) + ND^2 + D^3\right)$
RandNE [53]	$\mathcal{O}\left(ND^2 + \lvert E\rvert Dq\right)$
LouvainNE [51]	$\mathcal{O}\left(\lvert E\rvert\mathcal{H} + ND\right)$
ProNE [50]	$\mathcal{O}\left(ND^2 + \lvert E\rvert c\right)$
Verse [130]	$\mathcal{O}\left(N(\mathcal{W} + kD)\right)$
HBDM(-Re)	$\mathcal{O}\left(\alpha_2 N \log\left(N\right)D\right)$

in such a space. We here, also aim to define community-aware representations, meaning that each embedding should convey information about the community structure. Overall, we would like to define a Graph Representation Learning method expressing the desired properties of homophily and transitivity coupled with latent structure characterization and under a unique optimization procedure (i.e. no post-processing steps such as clustering procedures). Under such a direction, we will focus on finding mappings of the nodes into the unit D-simplex set, $\Delta^D \subset \mathbb{R}_+^{D+1}$ formally defined as

$$\Delta^D = \left\{ (x_0, \ldots, x_D) \in \mathbb{R}^{D+1} \,\middle|\, \sum_{d=0}^{D} x_d = 1,\ x_d \geq 0,\ \forall\, d \in \{0, \ldots, D\} \right\}.$$

In addition, we provide the standard 2-simplex in Figure 2.6 with an example of an embedding $\mathbf{z}_i$. A direct consequence of constraining node representation on the simplex is that the extracted node embeddings can convey information about latent community memberships. Numerous GRL methods lack guarantees for identifiable or unique solutions, making their interpretation heavily reliant on the initial setting of hyper-parameters. In this chapter, we are also focusing on the issue of identifiability. We aim to find identifiable solutions, though these can only be realized to the extent of permutation invariance, as described in Def. 1.

Definition 1 (**Identifiabilty**). *An embedding matrix* $\mathbf{Z}$ *whose rows indicate the corresponding node representations is called an identifiable solution up to a permutation if it holds* $\widetilde{\mathbf{Z}} = \mathbf{Z}\mathbf{P}$ *for a permutation* $\mathbf{P}$ *and a solution* $\widetilde{\mathbf{Z}} \neq \mathbf{Z}$.

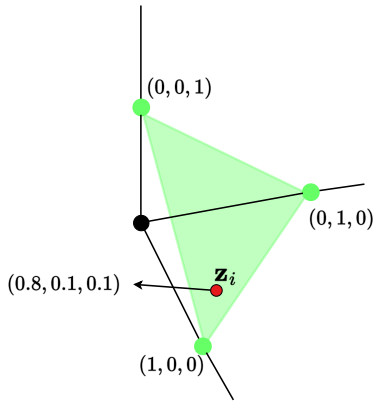
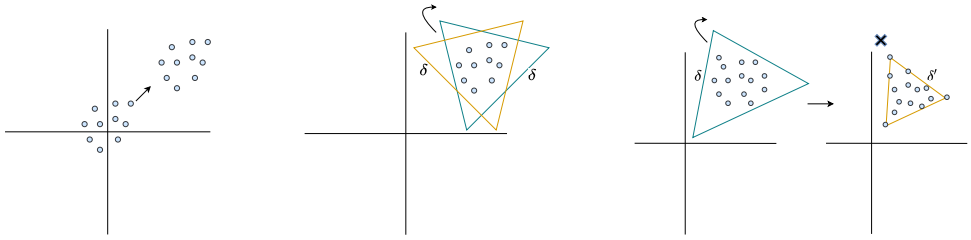


Figure 2.6. The standard 2-simplex in $\mathbb{R}^3$ which is a triangle. Any point $\mathbf{z}_i$ of the simplex lies on the affine hyperplane and is denoted with the green-colored area, and can be expressed as a convex combination of the three corresponding vertices (corners).



(a) Translation invariances.

(b) Rotation invariances.

(c) Decreased simplex volume ensuring identifiability.

Figure 2.7. A 2-dimensional latent space with the 2-simplex given as the green and yellow triangles, the blue points denote embedding positions of the LDM and δ is the simplex size [74].

2.5.1 Hybrid memberships under a latent distance model

We will consider a Poisson LDM, defining a log-likelihood over the adjacency matrix $\mathbf{Y}$ of the network as introduced by Equation (2.3). To combine powerful and community-aware representations, we propose the Hybrid-Membership Latent Distance Model (HM-LDM) with a log-rate based on the ℓ^2-norm as:

$$\log \lambda_{ij} = \Big(\gamma_i + \gamma_j - \delta^p \cdot ||\mathbf{z}_i - \mathbf{z}_j||_2^p\Big), \tag{2.12}$$

where $\mathbf{z_i} \in [0,1]^{D+1}$ are the latent embeddings constrained to the $D-$simplex, i.e. $\sum_{d=1}^{D+1} w_{id} = 1$, $\delta \in \mathbb{R}_+$ is the non-negative value controlling the simplex volume, and $\gamma_i \in \mathbb{R}$ a bias term of node $i \in \mathcal{V}$ accounting for node-specific effects such as degree heterogeneity. Lastly, p is the power of the ℓ_2 norm with $p \in \{1, 2\}$ which governs

the model specification. Specifically, power p modifies the effect of the embedding distances within the rate functions. In other words, in Equation 2.12 we constrain the latent space to the $D-$simplex, and the simplex's edge lengths (1-faces) are scaled by the non-negative constant δ, controlling the length of the sides of the simplex, and consequently, the volume of the simplex itself.

A notable characteristic of Equation (2.12), is that it resembles a positive Eigenmodel with random effects: $\tilde{\gamma}_i + \tilde{\gamma}_j + (\tilde{\mathbf{w}}_i \mathbf{\Lambda} \tilde{\mathbf{w}}_j^\top)$ where $\mathbf{\Lambda}$ is a diagonal matrix having non-negative elements, i.e. $\tilde{\gamma}_i = \gamma_i - \delta^2 \cdot ||\mathbf{z}_i||_2^2$, $\tilde{\gamma}_j = \gamma_j - \delta^2 \cdot ||\mathbf{z}_j||_2^2$ and $\tilde{\mathbf{z}}_i \mathbf{\Lambda} \tilde{\mathbf{z}}_j^\top = 2\delta^2 \cdot \mathbf{z}_i \mathbf{z}_j^\top$. Therefore, the squared Euclidean distance acts as a bridge between the traditional LDM and the non-negativity-constrained Eigenmodel. While not entirely conforming to the definition of a metric, the squared Euclidean distance still conveys homophily, resulting in an interpretable latent space. Although it doesn't precisely satisfy the triangle inequality, it maintains the order of pairwise Euclidean distances and is often favored in applications due to its nature as a strictly convex smooth function. By the well-known cosine formula, we have

$$||\mathbf{z}_i - \mathbf{z}_j||_2^2 = ||\mathbf{z}_i - \mathbf{z}_k||_2^2 + ||\mathbf{z}_k - \mathbf{z}_j||_2^2 - 2||\mathbf{z}_i - \mathbf{z}_k||_2 ||\mathbf{z}_k - \mathbf{z}_j||_2 \cos(\theta),$$

where $\theta \in (-\pi/2, \pi/2)$ represents the angle between $\mathbf{z}_i - \mathbf{z}_k$ and $\mathbf{z}_k - \mathbf{z}_j$. Note that the third term also approaches to 0 for $\theta \to \pi/2$. For the case where $\theta \in [\pi/2, 3\pi/2]$, it satisfies the triangle inequality: $||\mathbf{z}_i - \mathbf{z}_j||_2^2 \leq ||\mathbf{z}_i - \mathbf{z}_k||_2^2 + ||\mathbf{z}_k - \mathbf{z}_j||_2^2$.

The embedding vectors, $\{\mathbf{z}_i\}_{i=1}^N$ in Equation (2.12), are constrained to non-negative values and to sum to one. As a result, they are positioned on a simplex that shows the participation of node $i \in \mathcal{V}$ across $D+1$ latent communities. Any LDM can be constrained to the non-negative orthant without diminishing its performance or expressive power. Non-negative embeddings do not change the distance metric, as it remains constant under translation, as illustrated in Figure 2.7 (a). Furthermore, the D-dimensional non-negative orthant can be reconstructed by a large enough D-simplex. From these considerations, it can be effortlessly shown that for high values of the δ parameter in Equation (2.12), the sum-to-one constraint on the embeddings $\mathbf{Z}$ results in an unconstrained LDM, since the distances are unbounded when $\delta \to +\infty$. In this scenario, the memberships defined by the rows of matrix $\mathbf{Z}$ cannot be uniquely identified due to the distance invariance of rotation, as depicted in Figure 2.7 (b).

Nonetheless, by reducing the volume of the simplex (which is the same as lowering δ), the D-dimensional space of LDM will eventually cease to fit within the D-simplex, forcing the nodes to begin occupying the corners of this reduced simplex. A node is referred to as a *champion* if its latent representation corresponds to a standard binary unit vector.

Definition 2 (**Community champion**). *A node for a latent community is called champion if it belongs to the community (simplex corner) while forming a binary unit vector.*

Champion nodes hold considerable importance for the model's identifiability. If each corner of the simplex has at least one node (champion), then the model's solution

is identifiable (subject to a permutation matrix) (as per Def. 1). This occurs because any random rotation no longer maintains the solution's invariance, as illustrated by Figure 2.7 (c). It is evident that the scalar, δ, controls important properties, such as identifiability and the type of community memberships, while also the expressive capability of the model. Specifically, an HM-LDM with a large value of δ is equivalent to an unconstrained LDM that includes high expressive capability but also a rotation invariant space. In contrast, small values of δ result in identifiable solutions and can ultimately drive hard cluster assignments. Therefore, with very low values of δ, nodes are exclusively positioned at the corners of the simplex. Lastly, we can also find regimes of values for δ that offer identifiable solutions, and mixed-memberships but also performance similar to LDM, defining a silver lining.

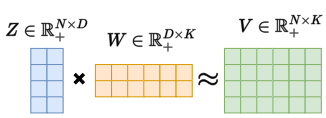
A different take on the identifiability of the model for $p = 2$, can also be given under the Non-negative Matrix Factorization (NMF) theory. Figures 2.8 (a) and (b), show both a non-symmetric and a symmetric NMF factorization. Specifically, a non-negative matrix V is factorized into two matrices Z and W, also non-negative. If the matrix is symmetric the factorization defines only the non-negative matrix Z. We will focus on undirected networks and make use of the symmetric NMF while we will not present extensions to bipartite and directed networks since they are trivial to obtain by switching to a non-symmetric NMF operation.

We can now easily show a re-parameterization of Equation (2.12) by $\tilde{\gamma}_i + \tilde{\gamma}_j + 2\delta^2 \cdot (\mathbf{z}_i \mathbf{z}_j^\top)$ as described in Equation (2.12). In such a formulation, the product $\mathbf{Z}\mathbf{Z}^\top$ defines a symmetric NMF problem which is an identifiable and unique factorization (up to permutation invariance) when $\mathbf{Z}$ is full-rank and at least one node resides solely in each simplex corner, ensuring separability [131, 132].

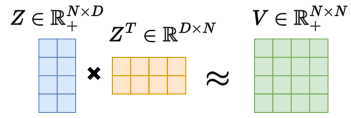
Under this NMF formulation, the product $\mathbf{z}_i \mathbf{z}_j^\top \in [0, 1]$ achieves its maximum value only when both nodes i and j reside in the same corner of the simplex. The parameter, δ, acts as a simple multiplicative factor in the first term of the objective function of HM-LDM, given in Equation (2.1), while in the second term acts as a power of the exponential function. For small values of δ, the model is biased towards hard latent community assignments of nodes since similar nodes achieve high rates only when they belong to the same latent community (simplex corner). On the other hand, nodes heading towards the simplex corners for large values of δ lead to an exponential change in the second term of the log-likelihood function given in Equation (2.1). Thus, a possible hard allocation of dissimilar nodes to the same community penalizes the likelihood severely. For this reason, high values of δ benefit mixed-membership allocations.

2.6 Signed integer weighted graphs

We continue now with the analysis of signed integer weighted graphs. Our aim is to learn representations for signed networks while expressing the properties of homophily, structure retrieval, and importantly heterophily/animosity as expressed by negative



(a) Non-Negative Matrix Factorization



(b) Symmetric Non-Negative Matrix Factorization

Figure 2.8. Factorization of a non-negative matrix V into two non-negative matrices Z and W. If the matrix is symmetric the factorization defines only the non-negative matrix Z.

relationships. Animosity or heterophily refers to the tendency for nodes to interact negatively when they express opposing or dissimilar views or opinions. In this setting, we would like to generalize transitivity properties to the expression of balance theory, a socio-psychological theory admitting four rules: "The friend of my friend is my friend", "The enemy of my friend is my enemy", "The friend of my enemy is my enemy", and "The enemy of my enemy is my friend", also presented in Figure 2.9. We can observe that transitivity is contained in balance theory and corresponds to the first case of Figure 2.9. We will move to the design of a GRL model able to characterize such properties and extend LDMs to the analysis of signed networks. In particular, we will utilize Archetypal Analysis [133, 134] allowing for model specifications allowing for archetype retrieval of relational data able to characterize network polarization.

2.7 The Skellam Latent Distance Model (SLDM)

We now generalize our main purpose which is to learn latent node representations $\{\mathbf{z}_i\}_{i\in\mathcal{V}} \in \mathbb{R}^D$ in a low dimensional space to signed networks $\mathcal{G} = (\mathcal{V}, \mathcal{Y})$ ($D \ll |\mathcal{V}|$). Therefore, the edge weights can take any integer value to represent the positive or negative tendencies between the corresponding nodes. We model these signed interactions among the nodes using the Skellam distribution [135], which can be formulated as the difference of two independent Poisson-distributed random variables ($y = N_1 - N_2 \in \mathbb{Z}$) with respect to the rates λ^+ and λ^-:

$$P(y|\lambda^+, \lambda^-) = e^{-(\lambda^+ + \lambda^-)} \left(\frac{\lambda^+}{\lambda^-}\right)^{y/2} \mathcal{I}_{|y|}\left(2\sqrt{\lambda^+\lambda^-}\right),$$

where $N_1 \sim Pois(\lambda^+)$ and $N_2 \sim Pois(\lambda^-)$, and $\mathcal{I}_{|y|}$ is the modified Bessel function of the first kind and order $|y|$. As far as we are aware, the Skellam distribution has not previously been used to model the likelihood of a network. We are introducing a novel latent space model that employs the Skellam distribution by adapting the latent distance model, originally devised for undirected and unsigned binary networks as a logistic regression model [76]. This was subsequently expanded to include various generalized linear models [117], such as the Poisson regression model tailored for

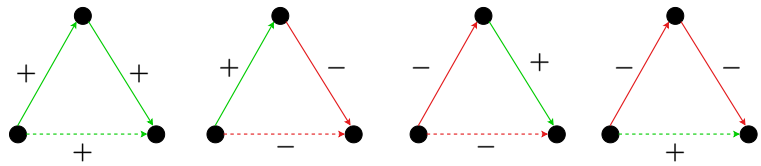


Figure 2.9. Graphical representation of the four balance theory properties, where black dots correspond to network nodes, green arrows as positive directed links, and red arrows as negative directed links. Dashed lines with arrows denote the inferred relationship under the balance theory. Analytically the panels show from left to the right, Case 1: The friend of my friend is my friend, Case 2: The enemy of my friend is my enemy, Case 3: The friend of my enemy is my enemy, Case 4: The enemy of my enemy is my friend.

integer-weighted networks. The negative log-likelihood of a latent distance model under the Skellam distribution can be formulated as follows:

$$\mathcal{L}(\mathcal{Y}) := \log p(y_{ij}|\lambda_{ij}^+, \lambda_{ij}^-) = \sum_{i<j} (\lambda_{ij}^+ + \lambda_{ij}^-) - \frac{y_{ij}}{2} \log \left(\frac{\lambda_{ij}^+}{\lambda_{ij}^-} \right) - \log\big(I_{ij}^*\big),$$

where $I_{ij}^* := \mathcal{I}_{|y_{ij}|} \left(2\sqrt{\lambda_{ij}^+ \lambda_{ij}^-} \right)$. As it can be noticed, the Skellam distribution has two rate parameters, and we consider them to learn latent node representations $\{\mathbf{z}_i\}_{i \in \mathcal{V}}$ by defining them as follows:

$$\lambda_{ij}^+ = \exp\big(\gamma_i + \gamma_j - ||\mathbf{z}_i - \mathbf{z}_j||_2\big), \tag{2.13}$$

$$\lambda_{ij}^- = \exp\big(\delta_i + \delta_j + ||\mathbf{z}_i - \mathbf{z}_j||_2\big), \tag{2.14}$$

where the set $\{\gamma_i, \delta_i\}_{i \in \mathcal{V}}$ denote the node-specific random effect terms, and $||\cdot||_2$ is the Euclidean distance function. More specifically, γ_i, γ_j represent the "social" effects/reach of a node and the tendency to form (as a receiver and as a sender, respectively) positive interactions, expressing positive degree heterogeneity (indicated by $+$ as a superscript of λ). In contrast, δ_i, δ_j provides the "anti-social" effect/reach of a node to form negative connections and thus models negative degree heterogeneity (indicated by $-$ as a superscript of λ). The rate formulation for the positive interactions λ_{ij}^+ naturally conveys the homophily property (a high positive rate is achieved when the distance is small) while negative interaction rate expression λ_{ij}^- models heterophily (a high negative rate is achieved when the distance is large). In addition, the corresponding rates in Equation (2.13) and Equation (2.14) satisfy balance theory, as it is a direct consequence of the high-order effects caused by the expression of homophily and heterophily, as seen by Figure 2.10.

By imposing standard normally distributed priors elementwise on all model parameters $\boldsymbol{\theta} = \{\boldsymbol{\gamma}, \boldsymbol{\delta}, \mathbf{Z}\}$, i.e., $\theta_i \sim \mathcal{N}(0, 1)$, We define a maximum a posteriori (MAP) estimation over the model parameters, via the loss function to be minimized (ignoring

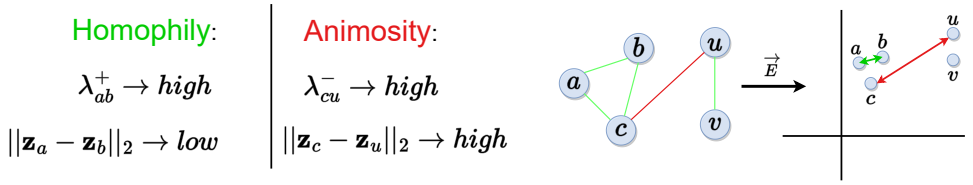


Figure 2.10. Expression of homophily and animosity as imposed by the Skellam Latent Distance model. Green lines correspond to positive interactions and red lines to negative interactions. Positively related nodes (i.e. pair $\{a, b\}$) are positioned close in space in order for the λ^+ rate to be high while negatively interacting nodes (i.e. pair $\{c, u\}$) are positioned far apart in space for the λ^- rate to be high.

constant terms):

$$Loss = \sum_{i<j} \left(\lambda^+_{ij} + \lambda^-_{ij} - \frac{y_{ij}}{2} \log \left(\frac{\lambda^+_{ij}}{\lambda^-_{ij}} \right) \right) - \sum_{i<j} \log I_{|y_{ij}|} \left(2\sqrt{\lambda^+_{ij}\lambda^-_{ij}} \right) + \frac{\rho}{2} \left(||\mathbf{Z}||^2_F + ||\boldsymbol{\gamma}||^2_F + ||\boldsymbol{\delta}||^2_F \right), \tag{2.15}$$

where $||\cdot||_F$ denotes the Frobenius norm. In addition, ρ is the regularization strength with $\rho = 1$ yielding the adopted normal prior with zero mean and unit variance. Importantly, by setting λ^+_{ij} and λ^-_{ij} based on Equation (2.13) and (2.14), the model effectively makes positive (weighted) links attract and negative (weighted links) deter nodes from being in proximity of each other.

2.7.1 Archetypal Analysis

Archetypal Analysis (AA) [133, 134] is a technique used in data clustering that identifies "archetypes" from a given observational dataset. An archetype is a pure or idealized form that represents an essential aspect or fundamental pattern within the data. In other words, archetypes are extreme points or corners of the convex hull of the data, and they can be thought of as the most representative or "extreme" examples of different behaviors or characteristics found in the data. In essence, AA is a powerful tool for clustering that provides a nuanced view of the data by identifying and utilizing extreme or "archetypal" patterns within the dataset. By expressing data in terms of these fundamental elements, AA offers an insightful perspective on the underlying structure and relationships within the data, aiding in cluster identification and interpretation. The definition of the embedded data points is given as follows:

$$\mathbf{X} \approx \mathbf{XCZ} \quad \text{s.t. } \boldsymbol{c}_d \in \Delta^N \text{ and } \mathbf{z}_j \in \Delta^D. \tag{2.16}$$

The archetypes, represented by the columns of $\mathbf{A} = \mathbf{XC}$, define the corners of the extracted polytope, serving as convex combinations of the observations. Mean-

while, $\mathbf{Z}$ outlines how each observation is reassembled as convex combinations of these extracted archetypes.

While Archetypal Analysis confines the representation within the convex hull of the data, alternative methods for modeling pure or ideal forms have included Minimal Volume (MV) approaches. One advantage of these approaches is that, unlike AA, they don't necessitate the existence of pure observations in the data. However, they come with the disadvantage of needing careful regularization tuning to determine an appropriate volume [136]. Additionally, the precise calculation of the volume for general polytopes demands the computation of determinants of the sum of all simplices that define the polytope [137] which comes with a high computational cost.

Archetypal Analysis and Minimal Volume extraction techniques have been recognized for their ability to uncover latent polytopes that define trade-offs. Within these polytopes, the vertices symbolize the maximally enriched and distinct aspects, or archetypes, which allow for the identification of specific tasks or prominent roles that the vertices represent [138, 139]. Owing to the computational challenges associated with regularizing high-dimensional volumes and the intricate fine-tuning required for those regularization parameters, our current focus is centered on polytope extraction as framed by the AA formulation, rather than adopting an MV approach.

2.7.2 A Generative Model of Polarization

Combining AA with the Skellam Latent Distance Model, i.e. constraining the latent space into a polytope allows for the modeling of polarization, as present in many signed networks. Specifically, we extend the Skellam LDM and express polarization based on defining node positions as convex combinations of the polytope - what we denote as a sociotope. The corners of the sociotope are considered the different "poles" that drive polarization and are sufficient to express the social dynamics of the network. Essentially, these uncovered "poles" are the extracted archetypes/extreme profiles as proposed by AA while every other node representation is a convex combination of these extremes.

In our generative model of polarization, we further suppose that the bias terms introduced in the definitions of the Poisson rates, $(\lambda_{ij}^{+}, \lambda_{ij}^{-})$, are normally distributed. Since latent representations $\{\mathbf{z}_i\}_{i\in\mathcal{V}}$ according to AA and MV lie in the standard simplex set Δ^D, we further assume that they follow a Dirichlet distribution. Formally,

we can summarize the generative model as follows:

$$
\begin{aligned}
\gamma_i &\sim \mathcal{N}(\mu_\gamma, \sigma_\gamma^2) && \forall i \in \mathcal{V}, \\
\delta_i &\sim \mathcal{N}(\mu_\delta, \sigma_\delta^2) && \forall i \in \mathcal{V}, \\
\mathbf{a}_d &\sim \mathcal{N}(\boldsymbol{\mu}_A, \sigma_A^2 \mathbf{I}) && \forall d \in \{1, \dots, D\}, \\
\mathbf{z}_i &\sim Dir(\boldsymbol{\alpha}) && \forall i \in \mathcal{V}, \\
\lambda_{ij}^+ &= \exp\big(\gamma_i + \gamma_j - \|\mathbf{A}(\mathbf{z}_i - \mathbf{z}_j)\|_2\big), \\
\lambda_{ij}^- &= \exp\big(\delta_i + \delta_j + \|\mathbf{A}(\mathbf{z}_i - \mathbf{z}_j)\|_2\big), \\
y_{ij} &\sim Skellam(\lambda_{ij}^+, \lambda_{ij}^-) && \forall (i,j) \in \mathcal{V}^2.
\end{aligned}
$$

According to the above generative process, positive ($\boldsymbol{\gamma}$) and negative ($\boldsymbol{\delta}$) random effects for the nodes are first drawn, upon which the location of extreme positions $\mathbf{A}$ (i.e., corners of the polytope denoted archetypes) are generated. In addition, as the dimensionality of the latent space increases linearly with the number of archetypes, i.e. $\boldsymbol{A}$ is a square matrix, with probability zero archetypes will be placed in the interior of the convex hull of the other archetypes. Subsequently, the node-specific convex combinations $\mathbf{Z}$ of the generated archetypes are drawn, and finally, the weighted signed link is generated according to the node-specific biases and distances between dyads within the polytope utilizing the Skellam distribution. The polarization level of the generative process can easily be controlled by the concentration parameter α of the Dirichlet distribution, defining the reconstruction matrix $\boldsymbol{Z}$.

2.7.3 The Signed Relational Latent Distance Model

For inference, we exploit how polytopes can be efficiently extracted using archetypal analysis. We, therefore, define the Signed Latent relational dIstance Model (SLIM) by defining a relational archetypal analysis approach endowing the generative model a parameterization akin to archetypal analysis in order to efficiently extract polytopes from relational data defined by signed weighted networks. Specifically, we formulate the relational AA in the context of the family of LDMs, as:

$$
\begin{aligned}
\lambda_{ij}^+ &= \exp\big(\gamma_i + \gamma_j - \|\mathbf{A}(\mathbf{z}_i - \mathbf{z}_j)\|_2\big) && (2.17) \\
&= \exp\big(\gamma_i + \gamma_j - \|\mathbf{RZC}(\mathbf{z}_i - \mathbf{z}_j)\|_2\big). && (2.18) \\
\lambda_{ij}^- &= \exp\big(\delta_i + \delta_j + \|\mathbf{A}(\mathbf{z}_i - \mathbf{z}_j)\|_2\big) && (2.19) \\
&= \exp\big(\delta_i + \delta_j + \|\mathbf{RZC}(\mathbf{z}_i - \mathbf{z}_j)\|_2\big). && (2.20)
\end{aligned}
$$

Notably, in the AA formulation $\mathbf{X} = \mathbf{RZ}$ corresponds to observations formed by convex combinations $\mathbf{Z}$ of positions given by the columns of $\boldsymbol{R}^{D \times D}$. Furthermore, in order to ensure what is used to define archetypes $\mathbf{A} = \mathbf{XC} = \mathbf{RZC}$ corresponds to observations using these archetypes in their reconstruction $\mathbf{Z}$, we define $\boldsymbol{C} \in \boldsymbol{R}^{N \times D}$

as a gated version of $\mathbf{Z}$ normalized to the simplex such that $\boldsymbol{c}_d \in \Delta^N$ by defining

$$c_{nd} = \frac{(\mathbf{Z}^\top \circ [\sigma(\mathbf{G})]^\top)_{nd}}{\sum_{n'}(\mathbf{Z}^\top \circ [\sigma(\mathbf{G})]^\top)_{n'd}} \tag{2.21}$$

in which $\circ$ denotes the elementwise (Hadamard) product and $\sigma(\mathbf{G})$ defines the logistic sigmoid elementwise applied to the matrix $\boldsymbol{G}$. As a result, the extracted archetypes are ensured to correspond to the nodes assigned the archetype, whereas the location of the archetypes can be flexibly placed in space as defined by $\mathbf{R}$. By defining $\mathbf{z}_i = \text{softmax}(\tilde{\mathbf{z}}_i)$ we further ensure $\mathbf{z}_i \in \Delta^D$. Examples of two latent spaces where archetypes correspond and do not correspond to observations using these archetypes are displayed in Figure 2.11 with the gate function securing informative polytopes. Such a formulation is necessary since there is no guarantee that making the polytope matrix $\boldsymbol{A}$ a free parameter will lead to an informative latent space. This comes as a consequence of the fact that a large enough volume of the polytope $\boldsymbol{A}$ matrix can enclose an unconstrained LDM and avoid representing trade-offs that would force the nodes to use the corners.

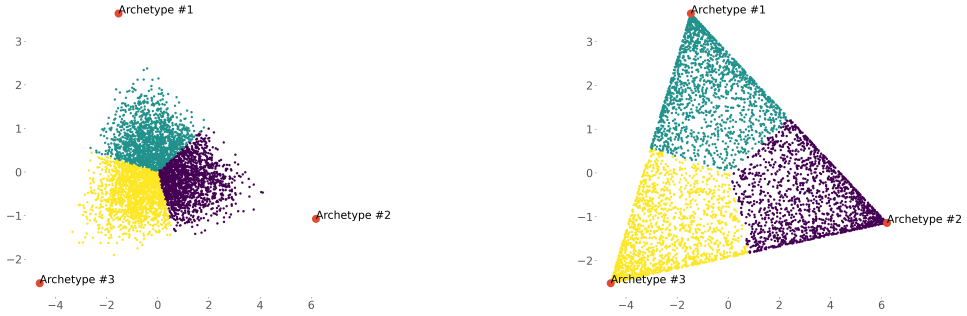
Importantly, the loss function of Equation (2.15) is adopted for the relational AA formulation forming the SLIM, with the prior regularization applied to the corners of the extracted polytope $\boldsymbol{A} = \mathbf{RZC}$ instead of the latent embeddings $\boldsymbol{Z}$ imposing a standard elementwise normal distribution as prior $a_{k,k'} \sim \mathcal{N}(0,1)$. Furthermore, we impose a uniform Dirichlet prior on the columns of $\boldsymbol{Z}$, i.e. $(\boldsymbol{z}_i \sim Dir(\mathbf{1}_D)$, this only contributes constant terms to the joint distribution, and therefore the maximum a posteriori (MAP) optimization only constant terms. As a result, the loss function optimized is given by Equation (2.15) replacing $\|\mathbf{Z}\|_F^2$ with $\|\boldsymbol{A}\|_F^2$.

2.8 Directed Case Model Formulations

As we have discussed, one of the most important properties of signed network models is the expression of balance theory which naturally describes directed relationships. In this section, we describe how our proposed frameworks can be extended to the study of directed networks (which at least for the SLIM formulations is not trivial). We further explore additional model formulations allowing for more capacity and expressive power.

2.8.1 The Skellam Latent Distance Model for the Directed Case (LDM)

Our main purpose here is to learn two latent node representations $\{\mathbf{z}_i\}_{i\in\mathcal{V}} \in \mathbb{R}^D$ and $\{\mathbf{w}_i\}_{i\in\mathcal{V}} \in \mathbb{R}^D$ in a low dimensional space for a given directed signed network $\mathcal{G} = (\mathcal{V}, \mathcal{Y})$ $(D \ll |\mathcal{V}|)$. The two sets of the latent embeddings correspond to modeling directed relationships $i \to j$ of nodes, with $\mathbf{z}_i$ the source node and $\mathbf{w}_j$ the target node,



(a) Non-informative polytope **(b)** Informative polytope

Figure 2.11. Example of two 2-dimensional polytopes projected into the first two principal components, defining a loose versus a tight latent space. Left panel: An example of a non-informative polytope where archetypes are not defined as observations belonging to the data, Right panel: Informative polytope where archetypes are defined as observations belonging to the data. Points are colored based on the archetype they express the maximum membership to.

and vice-versa for an oppositely directed relationship $i \leftarrow j$. Similar to the main paper, we can formulate the negative log-likelihood of a latent distance model under the Skellam distribution as:

$$\begin{aligned}\mathcal{L}(\mathcal{Y}) &:= \log p(y_{ij}|\lambda_{ij}^{+}, \lambda_{ij}^{-}) \\ &= \sum_{i,j} (\lambda_{ij}^{+} + \lambda_{ij}^{-}) - \frac{y_{ij}}{2} \log\left(\frac{\lambda_{ij}^{+}}{\lambda_{ij}^{-}}\right) - \log\left(\mathcal{I}_{|y_{ij}|}\left(2\sqrt{\lambda_{ij}^{+}\lambda_{ij}^{-}}\right)\right),\end{aligned}$$

For the directed case, the Skellam distribution has two rate parameters as well, and we consider them to learn latent node representations $\{\mathbf{z}_i\}_{i\in\mathcal{V}}$ and $\{\mathbf{w}_j\}_{j\in\mathcal{V}} \in \mathbb{R}^D$ by defining them as follows:

$$\lambda_{ij}^{+} = \exp\left(\beta_i + \gamma_j - ||\mathbf{z}_i - \mathbf{w}_j||_2\right), \tag{2.22}$$

$$\lambda_{ij}^{-} = \exp\left(\delta_i + \epsilon_j + ||\mathbf{z}_i - \mathbf{w}_j||_2\right), \tag{2.23}$$

where the set $\{\beta_i, \gamma_i, \delta_i, \epsilon_i\}_{i\in\mathcal{V}}$ denote the node-specific random effect terms. More specifically, the sender β_i and the receiver γ_j random effects represent the "social" reach of a node and the tendency to form positive interactions, expressing positive degree heterogeneity (indicated by $+$ as a superscript of λ). In contrast, δ_i and ϵ_j provide the "anti-social" sender and receiver effect of a node to form negative connections, and thus model negative degree heterogeneity (indicated by $-$ as a superscript of λ).

By imposing (as in the undirected case) standard normally distributed priors elementwise on all model parameters $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\epsilon}, \mathbf{Z}, \boldsymbol{W}\}$, i.e., $\theta_i \sim \mathcal{N}(0, 1)$, We

define a maximum a posteriori (MAP) estimation over the model parameters, via the loss function to be minimized (ignoring constant terms):

$$\begin{aligned} Loss = &\sum_{i,j}\left(\lambda_{ij}^{+}+\lambda_{ij}^{-}-\frac{y_{ij}}{2}\log\left(\frac{\lambda_{ij}^{+}}{\lambda_{ij}^{-}}\right)\right)-\sum_{i,j}\log I_{|y_{ij}|}\left(2\sqrt{\lambda_{ij}^{+}\lambda_{ij}^{-}}\right) \\ &+\frac{\rho}{2}\Big(||\mathbf{Z}||_F^2+||\mathbf{W}||_F^2+||\boldsymbol{\gamma}||_F^2+||\boldsymbol{\beta}||_F^2+||\boldsymbol{\delta}||_F^2+||\boldsymbol{\epsilon}||_F^2\Big), \end{aligned} \tag{2.24}$$

where $||\cdot||_F$ denotes the Frobenius norm. In addition, ρ is the regularization strength with $\rho = 1$ yielding the adopted normal prior with zero mean and unit variance.

2.8.2 The Signed Relational Latent Distance Model for Directed Networks

We formulate the relational AA in the context of the family of LDMs and for directed networks, as:

$$\lambda_{ij}^{+} = \exp\left(\beta_i+\gamma_j-\|\mathbf{A}(\mathbf{z}_i-\mathbf{w}_j)\|_2\right) \tag{2.25}$$

$$= \exp\left(\beta_i+\gamma_j-\|\mathbf{R}[\mathbf{Z};\mathbf{W}]\mathbf{C}(\mathbf{z}_i-\mathbf{w}_j)\|_2\right). \tag{2.26}$$

$$\lambda_{ij}^{-} = \exp\left(\delta_i+\epsilon_j+\|\mathbf{A}(\mathbf{z}_i-\mathbf{w}_j)\|_2\right) \tag{2.27}$$

$$= \exp\left(\delta_i+\epsilon_j+\|\mathbf{R}[\mathbf{Z};\mathbf{W}]\mathbf{C}(\mathbf{z}_i-\mathbf{w}_j)\|_2\right). \tag{2.28}$$

Notably, in the AA formulation for directed networks $\mathbf{X} = \mathbf{R}[\mathbf{Z};\mathbf{W}]$ corresponds to observations formed by the concatenations of the convex combinations $\mathbf{Z}$ and $\boldsymbol{W}$ of positions given by the columns of $\boldsymbol{R}^{D\times D}$. Similar to the undirected case, in order to ensure what is used to define archetypes $\mathbf{A} = \mathbf{XC} = \mathbf{R}[\mathbf{Z};\mathbf{W}]\mathbf{C}$ corresponds to observations using these archetypes in their reconstruction $[\mathbf{Z};\mathbf{W}]$, we define $\boldsymbol{C} \in \boldsymbol{R}^{2N\times D}$ as a gated version of $[\mathbf{Z};\mathbf{W}]$ normalized to the simplex such that $\boldsymbol{c}_d \in \Delta^{2N}$ by defining

$$c_{nd} = \frac{([\mathbf{Z};\mathbf{W}]^\top \circ [\sigma(\mathbf{G})]^\top)_{nd}}{\sum_{n'}([\mathbf{Z};\mathbf{W}]^\top \circ [\sigma(\mathbf{G})]^\top)_{n'd}}. \tag{2.29}$$

As a result, the extracted archetypes are ensured to correspond to the nodes assigned the archetype, whereas the location of the archetypes can be flexibly placed in space as defined by $\mathbf{R}$. By defining $\mathbf{z}_i = \text{softmax}(\tilde{\mathbf{z}}_i)$ and $\boldsymbol{w}_i = \text{softmax}(\tilde{\boldsymbol{w}}_i)$ we further ensure $\mathbf{z}_i, \boldsymbol{w}_i \in \Delta^K$.

As in the undirected case, the loss function of Equation (2.24) is adopted for the relational AA formulation forming the SLIM, with the prior regularization applied to the corners of the extracted polytope $\boldsymbol{A} = \mathbf{R}[\mathbf{Z};\mathbf{W}]\mathbf{C}$ instead of the latent embeddings $\boldsymbol{Z}, \boldsymbol{W}$ imposing a standard elementwise normal distribution as prior $a_{k,k'} \sim \mathcal{N}(0,1)$. Furthermore, we impose a uniform Dirichlet prior on the columns of $\boldsymbol{Z}, \boldsymbol{W}$, i.e. $(\boldsymbol{z}_i, \boldsymbol{w}_i \sim Dir(\mathbf{1}_K)$, this only contributes constant terms to the joint distribution. As a result, the loss function is given by Equation (2.24) replacing $\|\mathbf{Z}\|_F^2$ and $\|\boldsymbol{W}\|_F^2$ with $\|\boldsymbol{A}\|_F^2$ for the maximum a posteriori (MAP) optimization.

2.8.3 Model Extensions for Additional Capacity

Directed relationships usually require additional expressive capability than in the case of modeling unindicted relationships. For that, we will briefly discuss alternative model formulations, yielding different distances for the positive and negative rates to define additional expressive capability (as opposed to the standard model version where latent distances were shared across rates). We consider a formulation such as setting the Skellam rates as, $\lambda_{ij}^{+} = \exp(\beta_i + \gamma_j - ||\mathbf{z}_i - \mathbf{w}_j||_2)$ and $\lambda_{ij}^{-} = \exp(\delta_i + \epsilon_j - ||\boldsymbol{u}_i - \boldsymbol{w}_j||_2)$. Under this assumption, a positive directed relationship $(i \rightarrow j)$ shows that node i "likes" node j and "dislikes" node j if it is negative. The latent embedding $\boldsymbol{w}_j$ is then the receiver position for the "likes" and "dislikes" with embeddings $\mathbf{z}_i$ and $\boldsymbol{u}_i$ being the sender positions for positive and negative relationships, respectively. In this case, we introduce three latent embeddings instead of the conventional two for the undirected case. The disparity of location $\mathbf{z}_i$ and $\boldsymbol{u}_i$ here can point out how polarity is formed between the two regions of the latent space. This model specification introduces an additional regularization for the third embedding matrix $\mathbf{U}$ in the loss function of Equation (2.24). For the RAA case, we thereby define $\mathbf{X} = \mathbf{R}[\mathbf{Z}; \mathbf{U}; \mathbf{W}]$, i.e., as the concatenation of all three latent positions and with $\boldsymbol{C} \in \boldsymbol{R}^{3N \times D}$.

2.9 The Signed Hybrid-Membership Latent Distance Model

In the previous chapter, we extended LDMs to the study of signed networks while characterizing network polarization via the use of Archetypal Analysis and the Skellam distribution.

Whereas in SLIM the network representations were constrained to the convex hull as defined by the inferred representations, we briefly discussed additional modeling direction for the discovery of pure/ideal forms based on Minimal Volume (MV) approaches. More formally, such approaches can be defined as

$$\mathbf{X} \approx \mathbf{AZ} \quad \text{s.t. } vol(\mathbf{A}) = v \text{ and } \boldsymbol{z}_j \in \Delta^D, \tag{2.30}$$

where $\mathbf{A} \in \mathbb{R}^{(D+1)\times(D+1)}$ is the matrix describing the archetypes (extreme points of the convex hull) of the latent space, and $vol(\mathbf{A})$ is the volume of matrix $\mathbf{A}$ which can be expressed through the determinant as $|det(\mathbf{A})|$ when $\mathbf{A}$ is a square matrix [136,139]. A main advantage is that the extraction of distinct aspects/profiles through MV does not require the existence of "pure" observations defining the convex-hull or else the extracted polytope/simplex. As the volume decreases, observations are naturally "forced" to populate the corners of the polytope, yielding archetypal characterization when the reconstruction of data is defined through convex combinations of these corners.

The principal disadvantage of MV procedures lies in the meticulous requirement for regularization tuning to delineate volumes that both guarantee identifiability and retain sufficient capacity to represent the data with minimal reconstruction error [136]. Furthermore, analytical and tractable computation of the volume of polytopes requires calculating the sum of determinants for all simplexes used to construct the inferred polytope [137]. This is computationally expensive (especially in high dimensions) and sometimes unstable when $\mathbf{A}$ comes close to singular.

In this chapter, we constrain the columns of matrix $\mathbf{A}$ to the D-simplex with length δ. Thus, by controlling the volume of $\mathbf{A}$, we essentially define a constrained-to-simplexes MV approach. Calculating the volume for the $D-$simplex with length δ is straightforward and computationally efficient. Rather than including regularization over the volume of $\mathbf{A}$ in the loss function during inference, we deterministically control the simplex length δ which is given as an input to the model and is gradually decreased until uniqueness guarantees are obtained. Volume minimization can be obtained trivially by decreasing δ. Such a procedure gives us explicit control over the model capacity by fixing the volume which is harder to obtain with classical MV approaches where the volume expression is inserted in the loss function.

Essentially, by defining $\mathbf{A}$ as $\mathbf{A} = \delta \cdot \mathbf{I}$, with $\mathbf{I}$ being the $(D+1) \times (D+1)$ identity matrix, we obtain as a special case of archetypal analysis under a constrained MV formulation. In addition, if every corner of the introduced simplex is populated by at least one node champion we obtain unique representations defining hybrid memberships.

We now introduce the Signed Hybrid-Membership Latent Distance Model (sHM-LDM). The sHM-LDM is able to analyse signed networks, and similar to [75] it introduces two Skellam rate parameters as:

$$\lambda_{ij}^{+} = \exp\left(\beta_i + \beta_j - \delta^p ||\mathbf{z}_i - \mathbf{z}_j||_2^p\right), \tag{2.31}$$

$$\lambda_{ij}^{-} = \exp\left(\psi_i + \psi_j + \delta^p ||\mathbf{z}_i - \mathbf{z}_j||_2^p\right), \tag{2.32}$$

where again $\mathbf{z_i} \in [0,1]^{D+1}$ and $\sum_{d=1}^{D+1} z_{id} = 1$, $\delta \in \mathbb{R}_+$ and $\beta_i, \psi_j \in \mathbb{R}$ denote the node-specific random-effects. As explained in Section 2.7, β_i, β_j express positive degree heterogeneity while ψ_i, ψ_j models negative degree heterogeneity. The norm degree $p \in \{1,2\}$ controls the power of the ℓ^2-norm, and thus the model specification, as in the unsigned case.

As in [75], we define a maximum-a-posteriori (MAP) estimation, utilizing the Skellam likelihood over the adjacency matrix $\mathbf{Y}$ of the network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. We conditionally assume an independent likelihood given the unobserved latent positions and random effects. The corresponding loss function excluding constant terms is:

$$L = \sum_{i<j} \left(\lambda_{ij}^{+} + \lambda_{ij}^{-} - \frac{y_{ij}}{2} \log\left(\frac{\lambda_{ij}^{+}}{\lambda_{ij}^{-}}\right) \right) - \sum_{i<j} \log I_{|y_{ij}|}\left(2\sqrt{\lambda_{ij}^{+}\lambda_{ij}^{-}}\right) + \frac{\rho}{2}\left(||\boldsymbol{\beta}||_F^2 + ||\boldsymbol{\psi}||_F^2\right), \tag{2.33}$$

where $\mathcal{I}_{|y|}$ is the modified Bessel function of the first kind and order $|y|$, $||\cdot||_F$ denotes the Frobenius norm. In addition, ρ is the regularization strength where

$\rho = 1$ is assumed throughout this paper yielding a normal prior with zero mean and unit variance for the random effects. For the latent positions, we assume a uniform Dirichlet distribution as a prior which only adds a constant term in Equation 2.33 and thus is excluded.

Choosing the case where $p = 2$, meaning that the sHM-LDM utilizes the squared Euclidean norm, we are able once more to relate the model to an Eigenmodel by creating the following reparameterizations. For the rate responsible for positive interactions $\{\lambda_{ij}^{+}\}$ as: $\tilde{\beta}_i + \tilde{\beta}_j + (\tilde{\mathbf{w}}_i \mathbf{\Lambda} \tilde{\mathbf{w}}_j^{\top})$ where $\mathbf{\Lambda}$ is a diagonal matrix having non-negative elements, i.e. $\tilde{\beta}_i = \beta_i - \delta^2 \cdot ||\mathbf{w}_i||_2^2$, $\tilde{\beta}_j = \beta_j - \delta^2 \cdot ||\mathbf{w}_j||_2^2$ and $\tilde{\mathbf{w}}_i \mathbf{\Lambda} \tilde{\mathbf{w}}_j^{\top} = 2\delta^2 \cdot \mathbf{w}_i \mathbf{w}_j^{\top}$. Similarly, for the rate responsible for expressing animosity $\{\lambda_{ij}^{-}\}$ as: $\tilde{\psi}_i + \tilde{\psi}_j + (\tilde{\mathbf{w}}_i \mathbf{\Lambda} \tilde{\mathbf{w}}_j^{\top})$ where $\mathbf{\Lambda}$ is a diagonal matrix having non-positive elements, i.e. $\tilde{\psi}_i = \psi_i - \delta^2 \cdot ||\mathbf{w}_i||_2^2$, $\tilde{\psi}_j = \psi_j - \delta^2 \cdot ||\mathbf{w}_j||_2^2$ and $\tilde{\mathbf{w}}_i \mathbf{\Lambda} \tilde{\mathbf{w}}_j^{\top} = 2\delta^2 \cdot \mathbf{w}_i \mathbf{w}_j^{\top}$. We witness that homophily in the case of sHM-LDM is expressed through a non-negative Eigenmodel (as in the unsigned case) while animosity/heterophily is expressed through a non-positive Eigenmodel able to express stochastic equivalence [114]. These two formulations admit the same embedding matrix $\mathbf{W}$ which balances the expression of "opposing" forces (homophily and animosity) in the latent space. Lastly, for $p = 2$ both expressions admit to an NMF operation, obtaining an identifiable and unique factorization (up to permutation invariance) when $\mathbf{W}$ is full-rank and at least one node resides solely in each simplex corner [131, 132] as in the case of HM-LDM for unsigned networks.

2.10 Complexity analysis.

Modern graphs can potentially contain millions of nodes, with even billion-scale networks becoming already common in the real world. As a direct consequence, the computational scaling of GRL models is of vital importance. All of the proposed methods of this thesis, at their core, are distance models and thus they scale prohibitively as $\mathcal{O}(N^2)$ since the node pairwise distance matrix needs to be computed. This does not allow the analysis of large-scale networks. In Section 2.4, we showed how we can successfully scale LDMs while characterizing for structure at multiple scales, defining a linearithmic $\mathcal{O}(N \log N)$ space and time complexity. The HBDM methodology naturally extends to all of the proposed models in this chapter. Nevertheless, we chose to scale the rest of the models by adopting an unbiased estimation of the log-likelihood through random sampling [140] (it is an unbiased estimator since every node is sampled with an equal probability). More specifically, gradient steps are based on the log-likelihood of the block formed by a sampled (per iteration and with replacement) set S of network nodes, as:

$$\log P(Y|\boldsymbol{\lambda})) = \underbrace{\sum_{i<j:y_{ij}=1} \log(\lambda_{ij})}_{\text{Link Term } \mathcal{O}(S)} - \underbrace{\sum_{i<j} \lambda_{ij}}_{\text{Non-Link Term } \mathcal{O}(S^2)} \qquad \text{with } \; i, j \in S.$$

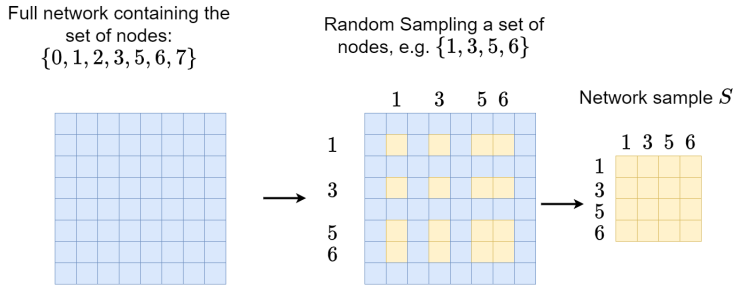


Figure 2.12. A random sampling procedure over an undirected network with eight total nodes and a sample size of four nodes. The node sample set defines a network block sample defining an $\mathcal{O}(S^2)$ space and time complexity.

This makes inference scalable defining an $\mathcal{O}(S^2)$ space and time complexity allowing for the analysis of large-scale networks. A toy example of a random sampling procedure is provided in Figure 2.12. More options for scalable inference of distance models have also been proposed in [141].

2.11 The Single-Event Poisson Process

In many studies, various real networks are represented with static structures, not taking advantage of the rich temporal information they may offer. In such a direction, researchers have considered the analysis of temporal networks both in discrete [83–87], as well as, continuous [69,88–91] time settings. Furthermore, important network types, with the prominent example of citation networks, are characterized by a temporal structure with links between a pair of nodes occurring maximum once throughout the time horizon. Such networks have traditionally been studied as static [4, 22, 37, 74, 99]. Contrary to such practices, we here introduce a framework utilizing a new likelihood formation under a Single-Event Poisson Process, capable of analyzing single-event networks, capitalizing on the rich temporal information that static models are blind to. In this regard, we assume that the studied temporal networks are composed maximally of a single event between a node pair (dyad), and once an event between two nodes has occurred no more events are admissible between these two nodes, see also Figure 1.1 (right panel).

Before presenting our modeling strategy for the links of networks, we will first establish the notations used throughout the sections referring to single-event networks. We utilize the conventional symbol, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, to denote a directed Singe-Event-Network over the timeline $[0, T]$ where $\mathcal{V} = \{1, \ldots, N\}$ is the vertex and $\mathcal{E} \subseteq \mathcal{V}^2 \times [0, T]$ is the edge set such that each node pair has at most one link. Hence, a tuple, $(i, j, t_{ij}) \in \mathcal{E}$, shows a directed event (i.e., instantaneous link) from source node j to target i at time $t_{ij} \in [T]$, and there can be at most one (i, j, t_{ij}) element for each

$(i, j) \in \mathcal{V}^{\in}$ and some $t_{ij} \in [0, T]$.

We always assume that the timeline starts at 0 and the last time point is T, and we represent the interval by symbol, $[T]$. We employ $t_1 \leq t_2 \leq \cdots \leq t_N$ to indicate the appearance times of the corresponding nodes $1, 2, \ldots, N \in \mathcal{V}$, and we suppose that node labels are sorted with respect to their incoming edge times. In other words, if $i < j$, then we know that there is a node $k \in \mathcal{V}$ such that $t_{ik} \leq t_{jl}$ for all $l \in \mathcal{V}$.

2.11.1 Inhomogenous Poisson Point Process

The inhomogeneous Poisson Point Process (IPP) has been a prominent method for modeling the number of events occurring between nodes at different times throughout the study period of the temporal network [142]. Such a process defines an event intensity yielding the Poisson process rate function which represents the average event density. The probability of sampling m event points in a time interval $[T]$ is given by as,

$$p_N(M(T) = m) := \frac{[\Lambda(0, T)]^m}{m!} \exp(-\Lambda(0, T)), \tag{2.34}$$

where $M(T)$ is the random variable showing the number of events occurring over the interval $[T]$, and $\Lambda(T) := \int_0^T \lambda(t')\mathrm{d}t'$ for the intensity function $\lambda : [T] \to \mathbb{R}^+$ (Please visit [143] for an overview). We point here once again, to earlier studies [99, 120] which have demonstrated that adopting the Poisson likelihood for modeling binary relationships does not degrade the methods' predictive performance.

We now focus on SENs and more specifically on citation networks, for which we employ an IPP for characterizing the occurrence time of a link (i.e., a single event point indicating the publication date and thus the citation time). This is unlike conventional practice in IPP literature which concentrates on modeling the occurrence of an arbitrary number of events between a pair of nodes. Consequently, we assume that a pair can have at most one edge (i.e., link), and we discretize the probability of sampling m events given in Equation (2.34) as having either one event or no event cases. More formally, by applying Bayes' rule, we can write it as a conditional distribution of $M(t)$ being equal to $m \in \{0, 1\}$ as follows:

$$\begin{aligned} p_{M|M\leq 1}\left(M(T) = m\right) &= \frac{p_{M,M\leq 1}\left(M(T) = m, M(T) \leq 1\right)}{p_{M\leq 1}\left(M(T) \leq 1\right)} \\ &= \frac{p_M\left(M(T) = m\right)}{p_M\left(M(T) = 0\right) + p_M\left(M(T) = 1\right)} \\ &= \frac{\exp\left(-\Lambda(T)\right)\left[\Lambda(T)\right]^m}{\exp(-\Lambda(T)) + \exp(-\Lambda(T))\Lambda(T)} \end{aligned} \tag{2.35}$$

The conditional probability of a single-event occurrence under the proposed *Single-Event Poisson Process* is given by:

$$p_{M|M\leq 1}\left(M(T)=1\right)=\frac{\Lambda(T)}{1+\Lambda(T)}. \tag{2.36}$$

Let now (Y,Θ) be a random variable where Y shows whether a link occurred and Θ indicates the time of the link occurrence. Then together with Eq (2.36), we can write the likelihood of (Y,Θ) evaluated at $(1,t^*)$ as follows:

$$\begin{aligned} p_{Y,\Theta}\left(1,t^*\right) &= p_Y\left\{Y=1\right\} p_{\Theta|Y}\left\{\Theta=t^*|Y=1\right\} \\ &= \left(\frac{\Lambda(T)}{1+\lambda(T)}\right)\left(\frac{\Lambda(t^*)}{\Lambda(T)}\right)=\frac{\lambda(t^*)}{1+\Lambda(T)} \end{aligned} \tag{2.37}$$

Consequently, the log-likelihood of the whole network, assuming that each dyad follows the Single-Event Poisson Process, can be written as:

$$\mathcal{L}_{SE-PP}(\Omega):=\log p(\mathcal{G}|\Omega)=\sum_{1\leq i,j\leq N}\Big(y_{ij}\log\lambda(t_{ij})-\log\left(1+\Lambda_{ij}(t_i,T)\right)\Big) \tag{2.38}$$

where Ω is the model hyper-parameters and $\Lambda_{ij}(t_i,T):=\int_{t_i}^{T}\lambda_{ij}(t')\mathrm{d}t'$. Note that for a homogenous Poisson process with constant intensity λ_{ij} for each node pair i and j, the probability of having an event throughout the timeline is equal to $\Lambda_{ij}(T)/(1+\Lambda_{ij}(T))=T\lambda_{ij}/(1+T\lambda_{ij})$ by Equation (2.36). In this regard, the objective function stated in Equation (2.38) is equivalent to a static Bernoulli model [76]:

$$\mathcal{L}_{Bern}(\Omega):=\log p(\mathcal{G}|\Omega)=\sum_{i,j\in\mathcal{V}}\Big(y_{ij}\log\left(\tilde{\lambda}_{ij}\right)-\log\left(1+\tilde{\lambda}_{ij}\right)\Big), \tag{2.39}$$

where we have used the re-parameterization $T\lambda_{ij}=\tilde{\lambda}_{ij}$.

2.12 Dynamic Impact Characterization

In the realm of impact analysis and risk assessment, characterizing dynamic events is pivotal in understanding and managing potential consequences. We know that papers generally undergo the process of aging over time since novel works introduce more original concepts. In this regard, we model the distribution of the impact of a paper $\{i\}$ by the TRUNCATED normal distribution:

$$f_i(t)=\frac{1}{\sigma}\frac{\phi(\frac{t-\mu}{\sigma})}{\Phi(\frac{\kappa-\mu}{\sigma})-\Phi(\frac{\rho-\mu}{\sigma})} \tag{2.40}$$

where μ and σ are the parameters of the distribution which lie in $(\rho,\kappa)\in\mathbb{R}$, $\phi(x)=\frac{1}{\sqrt{2\pi}}\exp\{(-\frac{1}{2}x^2)\}$, and $\Phi(\cdot)$ is the cumulative distribution function $\Phi(x)=$

$\frac{1}{2}\left(1+\text{erf}(\frac{x}{\sqrt{2}})\right)$. In addition, as an alternative impact function, and similar to [144], we consider the Log Normal distribution:

$$f_i(t) = \frac{1}{t\sigma\sqrt{2\pi}} \exp\left(-\frac{\ln(t-\mu)^2}{2\sigma^2}\right) \tag{2.41}$$

where μ and σ are the parameters of the distribution. Such distributions are particularly valuable for capturing the inherent variability and asymmetry in the lifecycle of a paper.

2.13 Single-Event Network Embedding by the Latent Distance Model

Our main purpose is to represent every node of a given single-event network in a low D-dimensional latent space ($D \ll N$) in which the pairwise distances in the embedding space should reflect various structural properties of the network, like homophily and transitivity [99]. For instance, in the *Latent Distance Model* [76], one of the pioneering works, the probability of a link between a pair of nodes depended on the log-odds expression, γ_{ij}, as $\alpha - \|\mathbf{z}_i - \mathbf{z}_j\|_2$ where $\{\mathbf{z}_i\}_{i\in\mathcal{V}}$ are the node embeddings, and $\alpha \in \mathbb{R}$ is the global bias term responsible for capturing the global information in the network. It has been proposed for undirected graphs but can be extended for directed networks as well by simply introducing another node representation vector $\{\mathbf{w}_i\}_{i\in\mathcal{V}}$ in order to differentiate the roles of the node as source (i.e., sender) and target (i.e., receiver). By the further inclusion of two sets of random effects $\{\alpha_i, \beta_j\}$ describing the in and out degree heterogeneity, respectively, we can define the log-odds expression as:

$$\gamma_{ij} = \alpha_i + \beta_j - \|\mathbf{z}_i - \mathbf{w}_j\|_2 \tag{2.42}$$

We can now combine a dynamic impact characterization function with the *Latent Distance Model*, to obtain an expression for the intensity function of the proposed *Single-Event Poisson Process*, as:

$$\lambda_{ij}(t_{ij}) = \frac{f_i(t_{ij}) \exp\{\alpha_i\} \exp\{\beta_j\}}{\exp\{\|\mathbf{z}_i - \mathbf{w}_j\|_2\}}. \tag{2.43}$$

Combining the intensity function of Equation (2.43) with the log-likelihood expression of Equation (2.38) yields the Dynamic Impact Single-Event Embedding Model (DISEE) model. Under such a formulation, we exploit the time information data indicating when links occur through time, so we can grasp a more detailed understanding of the evolution of networks, generate enriched node representations, and quantify a node's temporal impact on the network.

2.13.1 Case-Control Inference

With DISEE being a distance model, it scales prohibitively as $\mathcal{O}(N^2)$ since the all-pairs distance matrix needs to be calculated. In order to scale the analysis to large-scale networks we adopt an unbiased estimation of the log-likelihood similar to a case-control approach [141]. In our formulation, we calculate the log-likelihood as:

$$\begin{aligned}\log p_{ij}(\mathcal{G}|\Omega) &= \sum_{j:y_{ij}=1} \left(y_{ij} \log \left(\lambda_{ij}(t^*_{ij}) \right) - \log \left(1 + \int_{t_i}^{T} \lambda_{ij}(t')dt' \right) \right) \\ &+ \sum_{j:y_{ij}=0} - \log \left(1 + \int_{t_i}^{T} \lambda_{ij}(t')dt' \right) \\ &= l_1 + l_0 \end{aligned} \tag{2.44}$$

As already mentioned, large networks are usually sparse so the link (case) likelihood contribution term l_1 can be calculated analytically, even for massive networks. The non-link (control) likelihood contribution term l_0 has a quadratic complexity $\mathcal{O}(N^2)$ in terms of the size of the network, and thus its computation is infeasible. For that, we introduce an unbiased estimator for $l_{i,0}$ which is regarded as a population total statistic [141]. We estimate the non-link contribution of a node $\{i\}$ via:

$$l_{i,0} = \frac{N_{i,0}}{n_{i,0}} \sum_{k=1}^{n_{i,0}} - \log \left(1 + \int_{t_i}^{T} \lambda_{ik}(t')dt' \right), \tag{2.45}$$

where $N_{i,0}$ is the number of total non-links (controls) for node $\{i\}$, and $n_{i,0}$ is the number of samples to be used for the estimation. We set the number of samples based on the node degrees as $n_{i,0} = 5 * \text{degree}_i$. This makes inference scalable defining an $\mathcal{O}(cE)$ space and time complexity.

Part II

Graph Representation Learning of positive integer weighted networks

CHAPTER 3

A Hierarchical Block Distance Model for Ultra Low-Dimensional Graph Representations

Two of the most important properties that are found in graphs and especially in social networks are homophily and transitivity, as introduced in Chapter 2 of this thesis. The ultimate goal of Graph Representation Learning is to find mappings that define latent spaces where a graph is to be projected on. In such a space, closely related or connected nodes of a graph should be positioned in close proximity, in terms of their distance in the latent space. Additional properties of a successful and powerful method for GRL are firstly the scalability of the method, meaning that a model should offer competitive space and time computational complexities, and secondly, the ability to characterize the structure of networks that usually emerge at multiple scales. We here motivate our work, arguing that a model supporting such properties can potentially provide expressive and multi-purpose embeddings that can help us investigate the latent structures and perform downstream tasks on massive graphs.

For such a direction we turn to Latent Space Models, and more specifically to Latent Distance Models where the use of the Euclidean distance for the construction of the latent space of the network naturally conveys the main motivation of Graph Representation Learning. This comes as a direct consequence of the Euclidean metric choice, naturally representing homophily, transitivity, and high-order nodal proximity. Unfortunately, the latter two properties, of scalability and structure characterization, are not expressed by a classical Latent distance model, as it scales prohibitively as $\mathcal{O}(N^2)$ both in space and time complexity since it requires the computation of the

all-pairs Euclidean distance matrix. Infusing the Latent distance models with the ability to account for hierarchical representations, as well as, define complexities allowing for the analysis of modern and large-scale graphs will be the goal of this chapter. Importantly, we will exploit hierarchical representations in order to define a linearithmic (in terms of network nodes) space and time complexity, forming the Hierarchical Block Distance Model (HBDM).

3.1 Contributions

We reconcile hierarchical block structures emerging at multiple scales, scalability, and network properties such as homophily and transitivity through a new Graph Representation Learning approach, namely the Hierarchical Block Distance Model. This comes through the hierarchical approximation of the all-pairs Euclidean distance matrix that the LDM defines via a novel divisive Euclidean k-means algorithm. The procedure overview is provided in Figure 3.1. Analytically our contributions are outlined as:

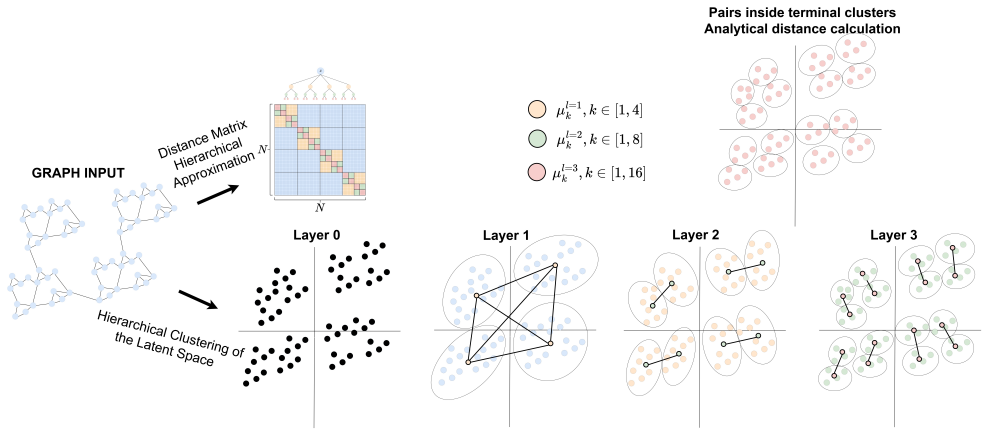


Figure 3.1. Hierarchical Block Distance Model procedure overview for a small network containing $N = 64$ number of nodes. Given a graph as input the model defines a divisive clustering of the latent space appropriate for a hierarchical approximation of the all-pairs distance matrix. Layer 1 defines the first divisive step, splitting the network embeddings into $K = \log N = \log 64 \approx 4$ clusters and the defined centroids $\boldsymbol{\mu}_k^{l=1}$ are used to approximate the node pairs belonging to different clusters (pairs inside the blue blocks of the displayed distance matrix). Then, binary splits are defined until each cluster contains a maximum of $\log N = \log 64 \approx 4$ points. Centroids of Layer 2 and 3 are used to approximate pair distances belonging only to the opposing cluster (the cluster that has the same parent cluster) as denoted with the yellow and green blocks of the displayed distance matrix. Distances for pairs inside the clusters of Layer 3 are calculated analytically and the clustering procedure terminates.

- We combine embedding and hierarchical characterizations for Graph Representation Learning, imposing a hierarchical block structure akin to stochastic block modeling (SBM) but explicitly accounting for homophily and transitivity properties throughout the inferred hierarchy.

- We design a hierarchical approximation of the the all-pairs Euclidean distance matrix admitting a linearithmic total time and space complexity, in terms of the number of nodes in the network (i.e., $\mathcal{O}(N \log N)$). Moreover, our proposed procedure is importance-aware meaning that the distance approximation becomes more accurate the more similar two nodes are.

- We define a novel objective function for an Euclidean k-means clustering algorithm, utilizing an auxiliary function framework to form an alternate least-squares convex optimization problem.

- We generalize the method for bipartite networks where multi-scale geometric representations, joint hierarchical structures, and community discovery are arduous tasks.

- We extensively and for the first time benchmark Latent Distance models against state-of-the-art GRL baselines and large-scale networks.

3.2 Experimental design, results, and key findings

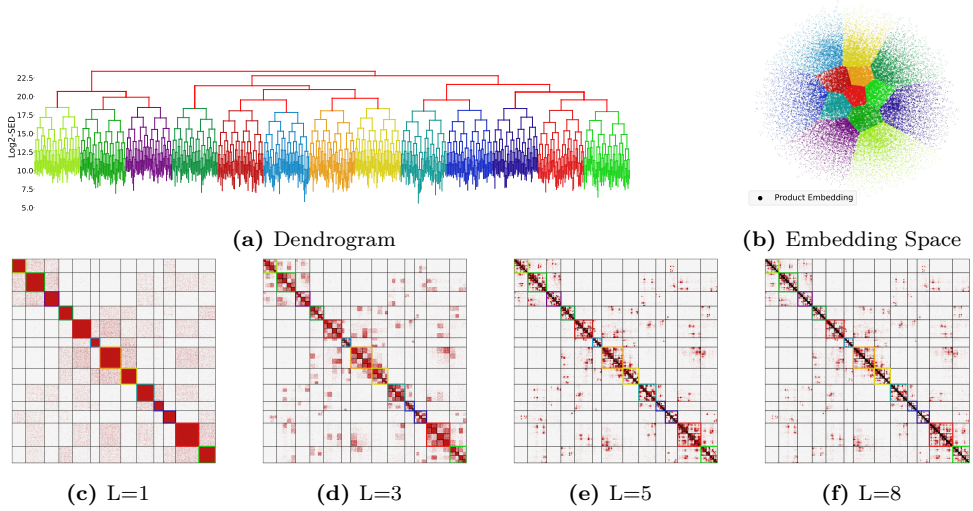
We adopt an extensive experimental evaluation framework that includes eleven prominent GRL and thirteen moderate-sized and large-scale networks, containing networks with more than one million nodes. We then establish the performance of our model in terms of multiple downstream tasks that include link prediction, node classification, network completion, and visualizations of both unipartite and bipartite networks. Furthermore, we make the downstream tasks setting even more challenging by constraining the embeddings to ultra-dimensions of a maximum of eight dimensions. Finally, we train the model by minimizing the negative log-likelihood via the Adam [145] optimizer.

The obtained results for the tasks of link prediction, network completion, and node classification showcase the favorable performance of HBDM against all baselines where in most cases the model significantly outperforms most baselines or defines on-par performance against the most competitive ones. Surprisingly, such a performance is achieved while using ultra-low-dimensional embeddings while we observe performance saturation when we reach $D = 8$ dimensions. Our results further highlight that the inferred hierarchical organization can facilitate accurate visualization of network structure even when using only $D = 2$ dimensional representations. Additionally, we show how our proposed framework extends the hierarchical multi-resolution structure to bipartite networks and provides the characterization of communities at

multiple scales, with superior performance in the task of link prediction. For two visualization examples, please visit Figure 3.2 where a product co-purchase unipartite Amazon network is provided, and Figure 3.3 where a Github user-product bipartite network is shown. An extensive study on the sensitivity of the three total hyperparameters of HBDM (dimension size, learning rate, and number of iterations) showed robust performance of the proposed frameworks on the downstream tasks. Importantly, the scalability of the model was studied both theoretically (in terms of the Big $\mathcal{O}$ notation) and experimentally, verifying the desired linearithmic space and time complexity. (For more details and the full experiment results please visit the full paper as provided in the Appendix in Section 9)

3.3 Conclusion

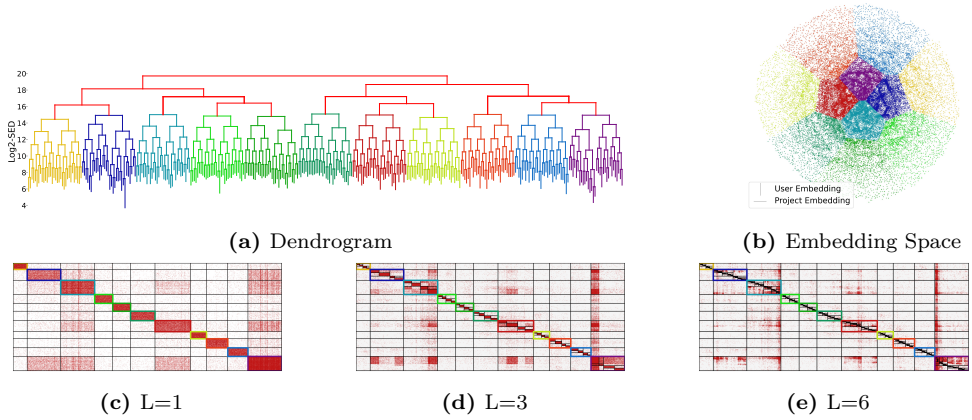
Overall, the use of an Euclidean distance metric for projecting complex networks into a latent space leads to high expressive capabilities even when using ultra-low dimensions. This allows for high network compression without a significant loss when performing multiple downstream tasks. The hierarchical approximation and extension of the LDM respected the so-desired properties of homophily and transitivity which allowed for high performance in downstream tasks. This came as an additional benefit to the scaling of LDM where we successfully exchanged a quadratic space and time complexity for a linearithmic one, allowing for scaling the analysis to large networks. The importance of accounting for multi-scale structures in complex networks was evident throughout the experiments where different resolution levels of the hierarchy led to different network communities and characterizations. All of these properties are generalized to bipartite networks successfully introducing multi-scale geometric representations, community discovery, and high downstream task performance even with ultra-low dimensions. Lastly, we considered multiple downstream tasks in each of which various baselines were found to be competitive against our HBDM frameworks. In general, the HBDM is characterized by the most consistent performance across tasks, making it state-of-the-art.



(a) Dendrogram **(b)** Embedding Space

(c) L=1 **(d)** L=3 **(e)** L=5 **(f)** L=8

Figure 3.2. *Amazon* network [146] dendrogram, embedding space and ordered adjacency matrices for the learned $D = 2$ embeddings of HBDM-Re and various levels (L) of the hierarchy [99].



(a) Dendrogram **(b)** Embedding Space

(c) L=1 **(d)** L=3 **(e)** L=6

Figure 3.3. *GitHub* network [147] dendrogram, embedding space and ordered adjacency matrices for the learned $D = 2$ embeddings of HBDM-Re and various levels (L) of the hierarchy [99].

CHAPTER 4

HM-LDM: A Hybrid-Membership Latent Distance Model

Community detection, alongside link prediction, and node classification, is one of the most notable downstream tasks in network science, and Graph Representation Learning. Often, graph embedding models are blind to community structures, or require additional post-processing steps (e.g. clustering procedures) to be able to account for community characterization. Furthermore, community detection can require soft, as well as, hard membership assignments to extract overlapping or non-overlapping communities, respectively. In the GRL literature, most algorithms impose hard community membership constraints with overlapping community detection (when possible) requiring careful designing and tuning of these models. Importantly, GRL models imposing overlapping community structures are usually able to be equally competitive to additional downstream tasks, such as link prediction, and node classification. Finally, many GRL approaches also do not provide identifiable or unique solution guarantees, so their interpretation highly depends on the initialization of the hyper-parameters, leading to the non-unique characterization of latent structures.

To provide a solution to such problems we here focus on combining a Non-Negative Matix Factorization with the Latent Distance Model. More specifically, we turn to the NMF theory and its uniqueness guarantees, under the scope of the LDM, where we can achieve unique soft and hard community memberships. Importantly, distance models offer high performance in additional important tasks, such as link prediction, and node classification, which is significantly superior to competing baselines in ultra-low dimensions [99]. As such, we aim to create an embedding model capable of characterizing community and latent structure without imposing any constraints on the type of memberships, providing unique representations but still explicitly accounting for homophily and transitivity, leading to superior performance on the main downstream tasks for ultra-low dimensions.

4.1 Contributions

Following the primary objective of modeling complex networks, we effectively learn graph representations in order to detect structures and predict link and node properties. In such a direction, we presently reconcile LSMs with latent community detection by constraining the LDM representation to the D-simplex forming the Hybrid-Membership Latent Distance Model (HM-LDM). Specifically, the HM-LDM offers part-based representations of networks relating to a Non-negative matrix factorization, while the LDM constructs low-dimensional latent spaces satisfying similarity properties such as homophily and transitivity. Additionally, we define a method that permits us to capture the latent community structure of the networks using a simple continuous optimization procedure over the log-likelihood of the network. Notably, unlike most existing approaches imposing hard community membership constraints, the assignment of community memberships in our proposed hybrid model can be controlled and altered through the simplex volume formed by the latent node representations. Specifically, we show that by systematically reducing the volume of the simplex, the model becomes unique and ultimately leads to hard assignments of nodes to simplex corners. We validate the effectiveness of the HM-LDM through extensive experiments, demonstrating accurate node representations and valid com-

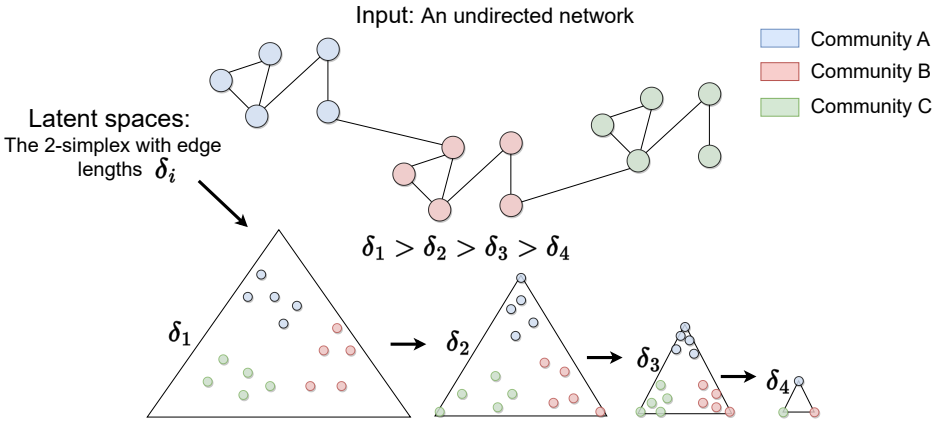


Figure 4.1. Hybrid Membership-Latent Distance Model procedure overview, considering a network with three communities and the 2-simplex. Given as an input an undirected network with a (latent) community structure decreasing the volume of the latent space starts characterizing the structure, defining initially mixed memberships while for a sufficiently shrunk volume, it defines hard assignments. Large simplex edge lengths (i.e. δ_1) define a large enough space that can enclose the whole representation without any decrease in the expressive capacity of the model. As the simplex edge lengths start being decreased more and more node representations move toward the corners (i.e. δ_1, δ_2), where eventually all node embeddings lie on a simplex corner (i.e. δ_4).

munity extraction in regimes that ensure identifiability. Importantly, we provide a systematic investigation of trade-offs between hard and mixed membership latent embeddings in terms of the model's ability to execute downstream tasks. We extensively evaluate the performance of the proposed method in link prediction, as well as, community discovery over various networks of different types, demonstrating that our model outperforms recent methods. The procedure overview is provided in Figure 4.1. Analytically our contributions are outlined as:

- We define community-aware latent representations by simply constraining the LDM to the D-simplex, forming the Hybrid-Membership Latent Distance Model which is explicitly accounting for homophily and transitivity properties, as well as, community and latent structure characterization.
- We design and empirically evaluate a continuous optimization procedure over the log-likelihood of the network by altering the latent space/simplex volume, allowing for control over soft and hard unique assignments to communities, and defining hybrid memberships.
- We provide uniqueness guarantees for the embeddings as obtained by the HM-LDM which is achieved up to permutation invariances.
- We show mathematically how a squared Euclidean LDM constrained to the D-simplex relates to the Non-Negative Matrix Factorization, defining a non-negative Latent Eigenmodel, and when such a factorization is unique.
- We systematically analyze the trade-offs that soft and hard community memberships define under the scope of link prediction and community detection tasks.
- We generalize the method for bipartite networks where structure-aware geometric representations, joint embedding spaces, and community discovery are arduous tasks.
- We extensively benchmark our proposed model against state-of-the-art GRL baselines, including models for both overlapping and non-overlapping community extraction under various and well-established network data.

4.2 Experimental design, results, and key findings

We adopt an extensive experimental evaluation framework that includes twelve prominent GRL baselines, including methods that use an NMF to learn their representation. In addition, we make use of four moderate-sized networks without known community labels to evaluate the model on link prediction and its ability to detect latent structures; four networks with known community labels to evaluate the model on its ability

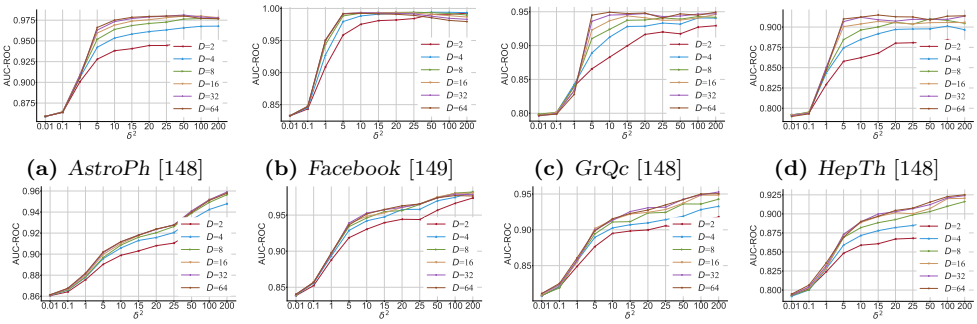
to perform community detection; and two bipartite networks to showcase the generalization of the model. We consider performance comparisons in downstream tasks of multiple HM-LDM versions defining big, moderate, and small latent space volumes, to understand the trade-offs that the resulting soft and hard memberships have on downstream tasks.

The obtained results for the link prediction task, showed that HM-LDM outperformed the considered baselines significantly, especially when compared to models that define mixed memberships under an NMF operation. For the community detection task, once more the HM-LDM outperformed or provided on par results with the most competitive baselines. For the study of trade-offs between soft and hard memberships or equivalently small and large volumes, the HM-LDM results showed high community detection results when the volume was particularly small and defined hard assignments to communities. Small volumes hampered the link prediction performance, as expected since such a small space decreases the expressive capability of the model. For large volumes, we saw a link prediction performance equivalent to the classical LDM since a large enough volume can absorb the whole non-negative orthant, making essentially the simplex constraint "powerless" as the latent distances can take very large values. These results are highlighted in Figure 4.2. The community detection results in high simplex volumes show a significant decrease in the performance as the model now suffers from identifiability issues. Importantly, the experiments for moderate-sized simplex volumes led to the existence of a silver lining where the model is identifiable and performed well on community detection while having almost an insignificant decrease in link prediction results when compared to the classical LDM. Identifiability results based on the type of community memberships for different simplex volumes are given in Figure 4.3 while latent community extraction examples on two real-world networks are provided in Figure 4.4. Finally, HM-LDM experiments on the two bipartite networks empirically showed successful latent structure extraction and identification. (For more details and the full experiment results please visit the full paper as provided in the Appendix in Section 9)

4.3 Conclusion

In this paper, we propose the HM-LDM, a model that reconciles network embedding and latent community detection. The approach utilizes normal and squared Euclidean specifications of the latent distance model. A squared Euclidean specification integrates the non-negativity-constrained Eigenmodel with the Latent Distance Model. We extensively showed that the model could be constrained to the simplex without losing expressive power. The reduced simplex provides unique representations, ultimately resulting in the hard clustering of nodes to communities when the simplex is sufficiently shrunk. Notably, the proposed HM-LDM combines network homophily and transitivity properties with latent community detection enabling explicit control of soft and hard assignment through the volume of the induced simplex. We observed

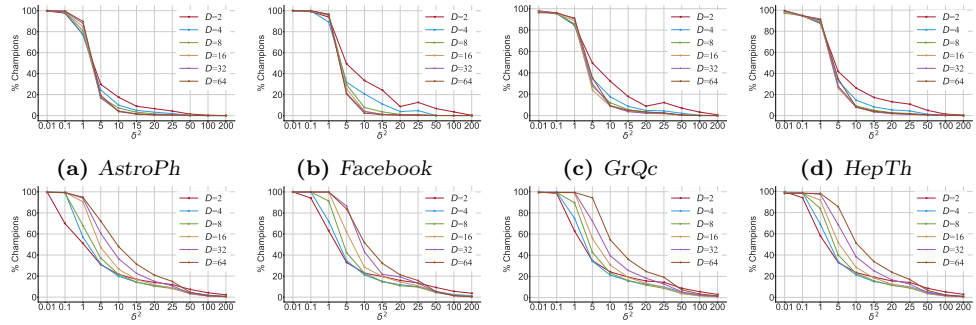
favorable link prediction performance in regimes in which the HM-LDM provides unique representations while enabling the ordering of the adjacency matrix in terms of prominent latent communities. Finally, we showed the ability of the model to extract valid community structures across multiple networks and showcased how the analysis extends to bipartite networks.



(a) *AstroPh* [148] **(b)** *Facebook* [149] **(c)** *GrQc* [148] **(d)** *HepTh* [148]

(e) *AstroPh* **(f)** *Facebook* **(g)** *GrQc* **(h)** *HepTh*

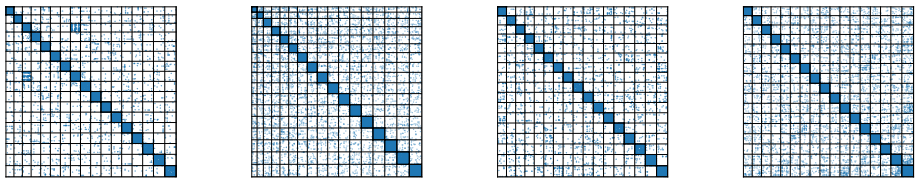
Figure 4.2. AUC-ROC scores as a function of δ^2 across dimensions for HM-LDM. Top row: $p = 2$. Bottom row $p = 1$ [74].



(a) *AstroPh* **(b)** *Facebook* **(c)** *GrQc* **(d)** *HepTh*

(e) *AstroPh* **(f)** *Facebook* **(g)** *GrQc* **(h)** *HepTh*

Figure 4.3. Total community champions (%) in terms of δ^2 across dimensions for HM-LDM. Top row: $p = 2$. Bottom row $p = 1$ [74].



(a) *GrQc* ($p = 2$) **(b)** *HepTh* ($p = 2$) **(c)** *GrQc* ($p = 1$) **(d)** *HepTh* ($p = 1$)

Figure 4.4. Ordered adjacency matrices based on the memberships of a $D = 16$ dimensional HM-LDM with δ values ensuring identifiability [74].

Part III

Graph Representation Learning of signed integer weighted networks

CHAPTER 5

Characterizing Polarization in Social Networks using the Signed Relational Latent Distance Model

Unlike traditional networks modeling only positive links or their absence between entities, signed networks can capture more complex relations, such as cooperative and antagonistic ties. They are instrumental in modeling more realistic and richer representations of real social structures. Hence, the analysis of signed networks can reveal significant insights into understanding how the network structure is actually formed. The proverb *"The enemy of my enemy is my friend"* is a very known example demonstrating that driving forces leading individuals to form connections are not merely positive inclinations. The *balance theory* explains these motives by proposing that individuals have an inner desire to provide balance and consistency in their relationships. Specifically, it is a socio-psychological theory admitting four rules: "The friend of my friend is my friend", "The enemy of my friend is my enemy", "The friend of my enemy is my enemy", and "The enemy of my enemy is my friend", also presented in Figure 2.9. In addition, signed networks can help us understand better ideological, as well as, affective polarization phenomena as present in social networks, as signed networks capture positive, negative, and neutral relationships between nodes and can characterize opposing views more accurately than unsigned networks. Ideological polarization refers to the substantial differences in how certain policies are viewed by elite or specialized groups, such as politicians, academics, or thought leaders. These groups might have widely divergent opinions on issues such as economic policies, social justice, or foreign relations. Essentially, ideological polarization is about the

"what" of political disagreements. Contrary to ideological polarization which focuses on differing opinions on policies, affective polarization refers to how ordinary voters feel about those differences, including strong emotions, such as anger or fear, that voters may feel towards the policy positions of parties or individuals they oppose. Affective polarization is more about "how" people feel about political disagreements rather than the content of the disagreements themselves. In addition, the media often presents the differences in policy positions in an extreme light, portraying them as existential or life-and-death threats. The culmination of these factors can lead to a divisive mentality ("us-versus-them") where different sides see each other not just as opponents with differing views but as existential threats. This binary view can stifle productive dialogue and compromise, leading to further polarization and possibly even hostility between different factions within society.

A first necessity to address and understand polarization phenomena is to devise a powerful framework for the analysis of signed networks. For that, we turn to the family of Latent Distance Models. Contrary to the case of unsigned networks, we now require on top of the homophily properties of a model to also be able to express animosity/heterophily in the latent space. This is necessary in order to extend the so-important transitivity properties present in the unsigned case with the more general balance theory. In addition, such a model should provide a valid likelihood function, describing both positive and negative interactions, as well as, defining a generative process over signed networks. In such a direction we turn to the Skellam distribution, a discrete probability distribution of the difference between two independent Poisson random variables, and extend Latent Distance Models forming the Skellam Latent Distance Model. In order to address and capture polarization phenomena we turn to Archetypal Analysis (AA) as introduced for observational data, and extend it to the analysis of relational data. Specifically, we focus on extreme positions and argue that the "us-versus-them" multipolarity, reinforced by homophily, animosity, and balance theory, can be represented by a latent position model. This model is applied to networks that are confined to a social space formed by a polytope akin to Archetypal Analysis, which we refer to as a "sociotope." The corners of the sociotope represent distinct aspects/poles formed by polarized network tendencies, where positive ties reinforce homophily among similar individuals, and negative ties repel dissimilar individuals to opposing poles. These multiple poles are important for defining the corners of the sociotope and revealing the different aspects of the social network. Thus for the modeling of signed networks and for the characterization of polarization, we present the Signed Relational Latent Distance Model, a combination of the Skellam Latent Distance Model and Archetypal Analysis.

5.1 Contributions

We extend Latent Distance Models to the analysis of signed networks, utilizing for the first time the Skellam distribution as network likelihood forming the Skellam

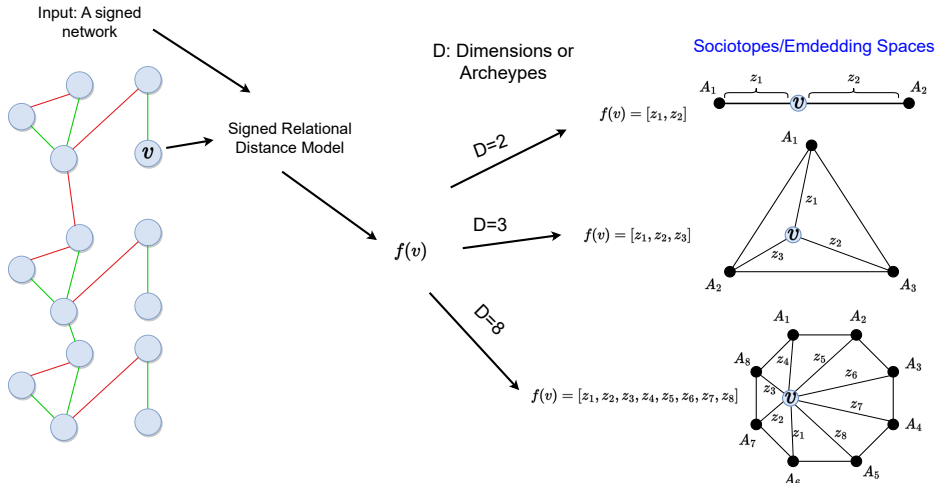


Figure 5.1. Signed Relational Latent Distance Model procedure overview. Analytically, for a given dimensionality and a signed network as inputs, the goal is to find a mapping function $f(\cdot)$ that projects a network node (e.g. $\{u\}$) into a latent space that is constrained to a polytope/sociotope, with every corner defining an archetype/extreme profile. Any node representation is characterized as a convex combination of the archetypes as these are the corner points of the convex hull defined by matrix $\boldsymbol{A}$. Sociotopes having dimensionality $D = 3, 8$ are denoted in a two-dimensional space for visualization purposes only.

Latent Distance Model. We show how such a model naturally conveys balance theory which comes as a direct consequence of the expression of homophily and heterophily properties our modeling design offers. We then reconcile Archetypal Analysis with the Skellam Latent Distance Model forming the Signed Relational Latent Distance Model and allowing for the characterization of polarization in terms of participation to extreme views or profiles as uncovered by the model. We extensively evaluate the performance of our frameworks and on four real social signed networks of polarization, we demonstrate that the models extract low-dimensional characterizations that well predict friendship and animosity while the Signed Relational Latent Distance Model provides interpretable visualizations defined by extreme positions when restricting the embedding space to polytopes akin to Archetypal Analysis. Furthermore, we successfully showcase a generative process allowing for the creation of networks with a controlled level of polarization while we further show how our frameworks generate accurate network representation when learning from real networks. The procedure overview is provided in Figure 5.1. Analytically our contributions are outlined as:

- We, for the first time, utilize the Skellam distribution as a network likelihood forming the Skellam Latent Distance model which satisfies the balance theory. We further, under the Skellam distribution, provide rate specifications allowing for different levels of model capacity, performance, latent space interpretation,

and the modeling of directed and undirected relationships.

- We present the Signed Relational Latent Distance Model, a novel method that extends Archetypal Analysis to relational data. We discuss how such a model successfully characterizes network polarization based on the discovery of distinct and extreme profiles being present in signed networks.
- We, contrary to the state-of-art, define generative models capable of generating signed networks of different polarization levels. Furthermore, we showcase the generative capabilities of our model on both real and artificial data, experimentally verifying that our model formulation can distinguish the different levels of network polarization.
- We extensively benchmark our proposed model against state-of-the-art GRL baselines designed for the analysis of signed networks. In multiple task settings, including sign link prediction, as well as, the more challenging task of signed link prediction, our model returns superior performance in most cases.
- We showcase how sociotope visualizations facilitate the characterization of network polarization, and importantly the successful discovery of influential nodes behaving as the driving forces of polarization for both directed and undirected settings.

5.2 Experimental design, results, and key findings

We employ four networks, describing electoral voting records and opinions. We benchmark the performance of our proposed frameworks against five prominent signed Graph Representation Learning methods, including random-walk-based methods and graph neural networks. We create a test set by removing 20% of the total network links while preserving connectivity on the residual network. We define two prediction tasks, *Link sign prediction (p@n):* In this setting, we utilize the link test set containing the negative/positive cases of removed connections. We then ask the models to predict the sign of the removed links. *Signed link prediction:* A more challenging task is to predict removed links against disconnected pairs of the network, as well as, infer the sign of each link correctly. For that, the test set is split into two subsets positive/disconnected and negative/disconnected. We then evaluate the performance of each model on those subsets. The tasks of signed link prediction between positive and zero samples are denoted as $p@z$ while the negative against zero is $n@z$. Furthermore, as the Signed Relational Latent Distance Model formulation facilitates the inference of a polytope describing the distinct aspects of networks, we visualize the latent space across $D = 8$ dimensions for all of the corresponding networks. To facilitate visualizations we use Principal Component Analysis (PCA), and project the space based on the first two principal components of the final embedding matrix. In addition, we provide circular plots where each archetype of the polytope is mapped

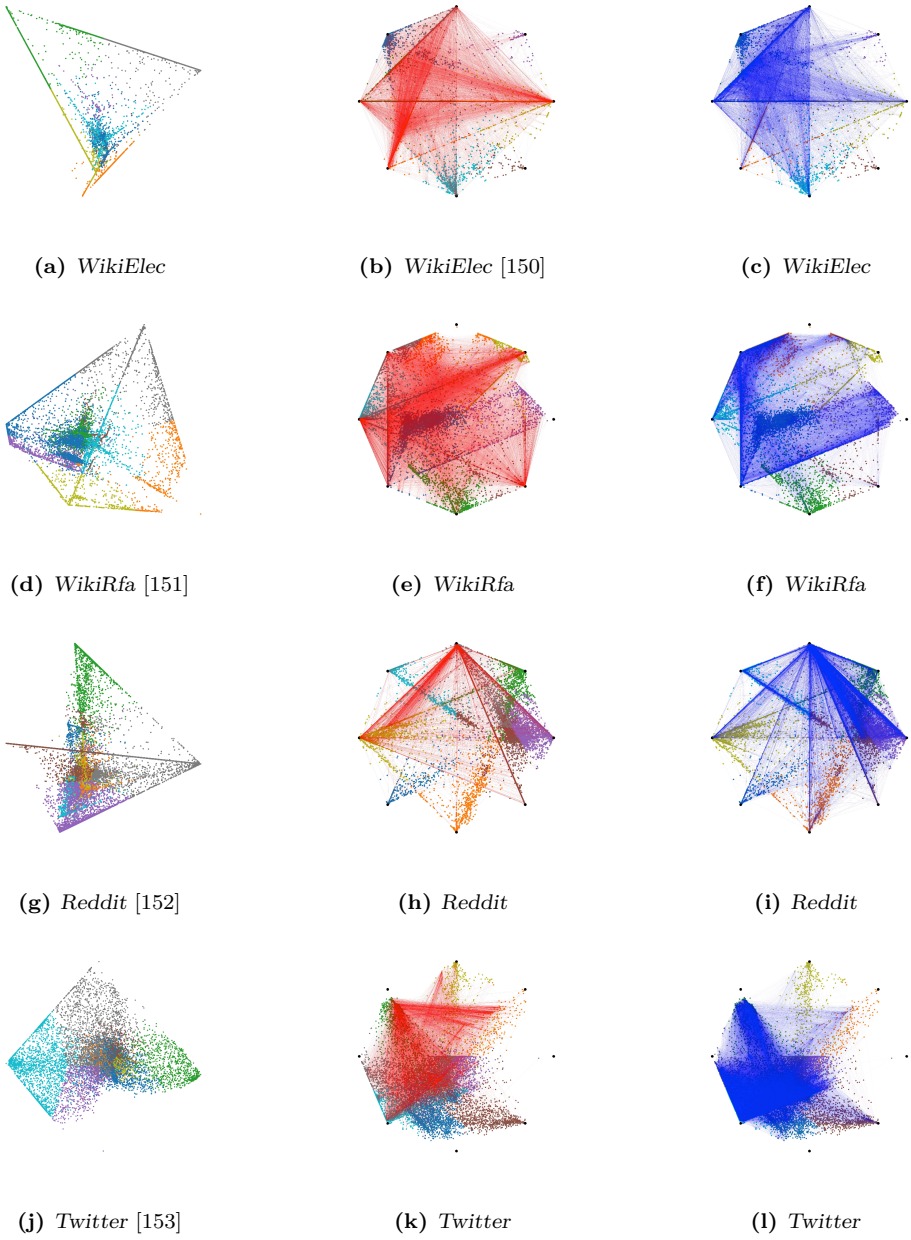
to a circle every $\text{rad}_d = \frac{2\pi}{D}$ radians, with D being the number of archetypes. Such polytope visualizations can be found in Figure 5.2.

During the evaluation, we focused on certain scoring metrics that are suitable for highly imbalanced data sets, specifically the AUC-ROC score and the AUC-PR score. We applied these scores to assess two particular tasks: link sign prediction and signed link prediction. In these tasks, we found that both the Skellam Latent Distance Model and the Signed Relational Latent Distance Model performed competitively when measured against all baseline models. Specifically, in most cases, our frameworks outperformed significantly most baselines or defined on-par performance against the most competitive ones. What further adds to the appeal of our models is that they are also designed with generative processes (for an example please visit Figure 5.3), making them particularly well-suited for the analysis of signed networks. By visualizing the sociotopes, we illustrate how the polytope method can successfully identify extreme positional nodes within the network. To put it more clearly, in all networks, there is at least one archetypal node that functions as a "dislike" hub, and at least one that operates as a "like" hub. These archetypes are characterized by having high values of either negative or positive interactions, respectively. In some networks, we also notice archetypes with a very low degree of connection. This phenomenon can be explained by the fact that some nodes, which are only associated with negative interactions, are pushed away from the main cluster of nodes. These isolated nodes can be considered "outliers" within the sociotope. However, such outliers are not merely anomalies but are discovered since they provide high expressive power for the model (allowing for a large volume of the polytope). (For more details and the full experiment results please visit the full paper as provided in the Appendix in Section 9)

5.3 Conclusion

The Skellam Latent Distance Model and Signed Latent Relational Distance Model that we have proposed allow for an easily interpretable visualization of signed networks, performing well in link-related prediction tasks when focusing on weighted signed networks. The Skellam Latent Distance Model extends the representation power of classical LDMs generalizing homophily and transitivity properties to the expression of balance theory. In addition, the Signed Latent Relational Distance Model defines a space that is constrained to polytopes, a feature that enables us to identify unique characteristics in social networks. This allows for the detection of extreme positions within the network, a process similar to traditional archetypal analysis but adapted for graph-structured data. The Skellam distribution is particularly useful for the modeling of signed networks, adding depth to our understanding of these structures; while the relational extension of Archetypal Analysis can be used for different likelihood specifications, under general Latent Distance Models. In summary, this study lays the groundwork for utilizing new likelihood formulations that

are suitable for analyzing weighted signed networks while it extends concepts similar to Archetypal Analysis to a broader context, offering a new way to analyze networks.



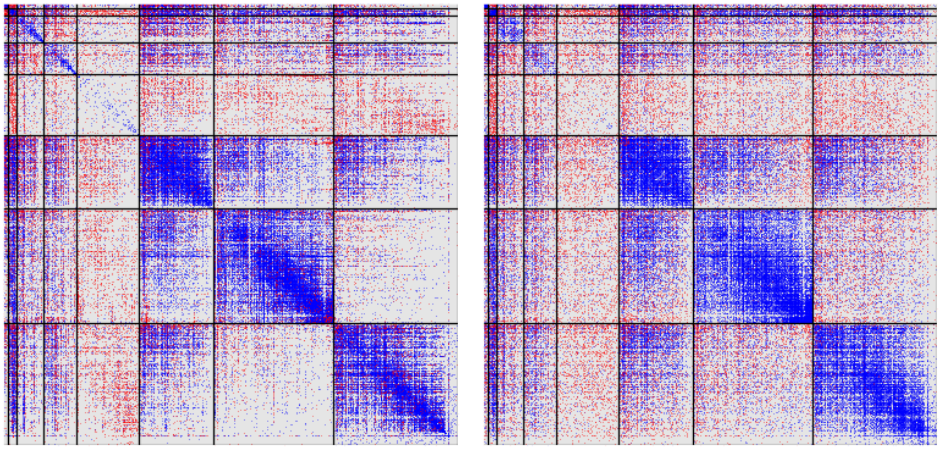
(a) *WikiElec* **(b)** *WikiElec* [150] **(c)** *WikiElec*

(d) *WikiRfa* [151] **(e)** *WikiRfa* **(f)** *WikiRfa*

(g) *Reddit* [152] **(h)** *Reddit* **(i)** *Reddit*

(j) *Twitter* [153] **(k)** *Twitter* **(l)** *Twitter*

Figure 5.2. Inferred polytope visualizations for various networks. The first column showcases the $K = 8$ dimensional sociotope projected on the first two principal components (PCA) — second and third columns provide circular plots of the sociotope enriched with the negative (red) and positive (blue) links, respectively [75].



(a) Ground Truth: $(.003, 78\%, 22\%)$ **(b)** Generated: $(.003, 76\%, 24\%)$

Figure 5.3. *wikiElec* ground truth (left) adjacency matrix and generated (right) adjacency matrix based on inferred parameters with a SLIM **without regularization priors** over the parameters. The parenthesis shows the network statistics as: (density,% of positive (blue) links,% of negative (red) links). All network adjacency matrices are ordered based on $\mathbf{z}_i$, in terms of maximum archetype membership and internally according to the magnitude of the corresponding archetype most used for their reconstruction [75].

CHAPTER 6

A Hybrid Membership Latent Distance Model for Signed Integer Weighted Networks

Signed networks, unlike traditional networks that model only positive and neutral connections, capture complex relations like cooperation and antagonism, providing a more realistic view of social structures. The balance theory with its four rules exemplifies the driving forces behind these connections, encompassing positive, negative, and neutral relationships. These concepts allow for a better understanding of phenomena such as ideological and affective polarization, which involve significant differences in how policies are viewed by various groups and the intense emotions voters may feel towards differing positions. Usually, signed networks contain nodes concentrating a high degree of both positive and negative ties. Such high-degree nodes act as driving forces of polarization, forming an archetypal ideology. This can be realized when considering the properties of balance theory, e.g. "The enemy of my friend is my enemy". In such cases, extreme profiles in networks can be easily uncovered by a model constraining the network projection into polytopes. Unfortunately, there is no guarantee that such "pure" nodes will be always present in polarized networks. For that, additional approaches have been developed such as Minimum Volume, where data representations are constrained to a polytope under a minimum volume constraint. As the volume decreases nodes are pushed to the corners of the defined space providing extreme profile characterizations akin to archetypal analysis. Such procedures are traditionally expressed by a high computational complexity since the volume calculation of a polytope requires calculating the sum of determinants for all simplexes used to construct the polytope which is particularly expensive, especially in high dimensions. Finally, many GRL approaches also do not provide identifiable or unique solution guarantees, so their interpretation highly depends on the initial-

ization of the hyper-parameters, leading to the non-unique characterization of latent structures

To derive an efficient Minumum Volume approach we turn to Latent Distance models and the Skellam distribution to form the Signed Hybrid-Membership Latent Distance Model. This new model, inspired by recent advances in Graph Representation Learning [75], is designed to highlight and uncover the unique characteristics of signed networks. Specifically, we constrain the latent space to the D-simplex. We show that the Signed Hybrid-Membership Latent Distance Model relates to archetypal analysis for relational data as a minimal volume approach and as a special case when polytopes are constrained to simplexes. Extraction of distinct aspects/profiles through MV does not require the presence of "pure" observations defining the convex-hull or else the extracted polytope/simplex. As the volume decreases, observations are "forced" to populate the corners of the polytope, yielding archetypal characterization when the reconstruction of data is defined through convex combinations of these corners. Based on the volume size we are able to control the type of memberships in these convex combinations. Specifically, we show that large volumes allow nodes to be expressed through many archetypes but as the volume decreases trade-offs are emerging, forcing nodes to collapse onto a unique archetype. Furthermore, constraining the polytope to the D-simplex allows for a trivial volume calculation which we can control simply by the edge length (1-faces) value of the simplex. We denote the edge length of the D-simplex as δ which is provided to the model as a continuously decreasing hyperparameter and as a consequence the model defines a continuously decreasing simplex volume, yielding archetypal characterization. Under such a formulation, we provide uniqueness guarantees by extending the Non-Negative Matrix Factorization theory to the study of signed networks which are achieved up to a permutation matrix.

6.1 Contributions

We presently derive a Minimum Volume approach for the archetypal characterization and analysis of signed networks forming the Signed Hybrid-Membership Latent Distance Model. Specifically, we show that constraining the polytope to the D-simplex comes with no loss of expressive power when the volume of the simplex is not significantly decreased. The model is characterized by high predictive performance in simplex volumes providing identifiable solutions and uncovering distinct aspects of signed networks. Importantly, by controlling the simplex volume we are able to control the type of memberships or participation across the different archetypes/pure forms. Decreasing sufficiently the volume of the latent space forces nodes to converge to their core ideologies/archetypes allowing for the expression of trade-offs. Furthermore, we consider different model specifications utilizing both the traditional Euclidean distance, as well as, the squared Euclidean distance. For the latter, we show that the Signed Hybrid-Membership Latent Distance Model is the combination

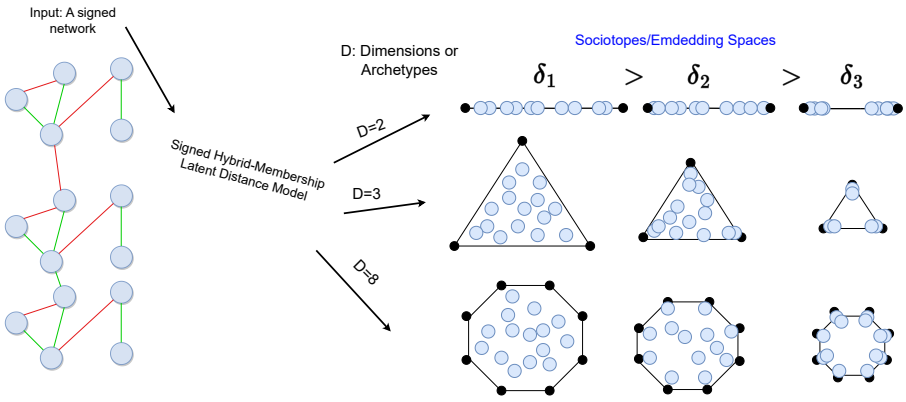


Figure 6.1. Signed Hybrid-Membership Latent Distance Model procedure overview. Analytically, for a given dimensionality D and a signed network as inputs, the model defines the $(D-1)$-simplex with edge length δ_i. As the length decreases the nodes start to populate the corners uncovering extreme profiles present in the graph data. This corresponds to the Archetypal Analysis of relational data when the polytope is constrained to the $(D-1)$-simplex and naturally extends hybrid memberships coupled with latent distance models to the analysis of signed networks.

of a non-negative Eigenmodel expressing homophily and a non-positive Eigenmodel yielding animosity/heterophily properties able to express stochastic equivalence. We benchmark the performance of our model against prominent signed network representation learning approaches and across four real signed networks, while we extend the analysis to two real bipartite networks. The procedure overview is provided in Figure 6.1. Analytically our contributions are outlined as:

- We, successfully derive a Minimum Volume approach for the analysis of signed networks, offering archetypal characterization. Constraining the polytope to the D-simplex, alleviates any computational burdens and restrictions that characterize the volume calculation of high-dimensional general polytopes.
- We design and empirically evaluate a continuous optimization procedure over the log-likelihood of the network by altering the latent space/simplex volume, allowing for control over the memberships across the different archetypes/pure forms.
- We provide uniqueness guarantees for the embedding solution as obtained by the Signed Hybrid-Membership Latent Distance Model which is achieved up to permutation invariances.
- We show mathematically how a squared Euclidean Skellam Latent Distance Model constrained to the D-simplex relates to the Non-Negative Matrix Factorization, defining a non-negative and a non-positive Latent Eigenmodels, and when such these factorizations are unique.

- We systematically analyze the trade-offs that soft and hard archetypal characterizations define under the scope of signed link prediction and sign link prediction tasks.
- We generalize the method for bipartite networks where archetypal-aware geometric representations, joint embedding spaces, and extreme node discovery are arduous tasks.
- We extensively benchmark our proposed model against state-of-the-art GRL baselines, including models utilizing graph neural networks and random walk approaches under various and well-established network data

6.2 Experimental design, results, and key findings

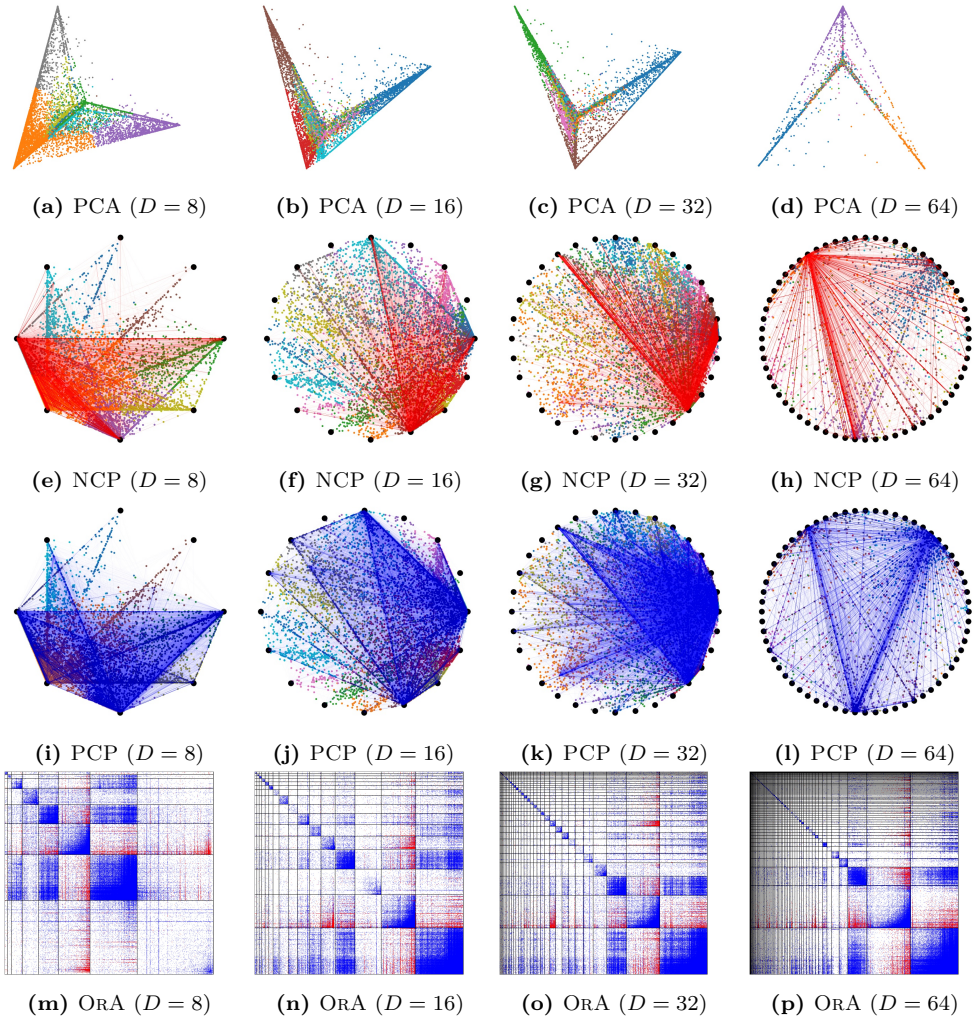
We employ four unipartite and two bipartite networks, describing electoral voting records and opinions. We benchmark the performance of our proposed frameworks against five prominent signed Graph Representation Learning methods, including random-walk-based methods and graph neural networks. We create a test set by removing 20% of the total network links while preserving connectivity on the residual network. We define two prediction tasks, *Link sign prediction (p@n):* In this setting, we utilize the link test set containing the negative/positive cases of removed connections. We then ask the models to predict the sign of the removed links. *Signed link prediction:* A more challenging task is to predict removed links against disconnected pairs of the network, as well as, infer the sign of each link correctly. For that, the test set is split into two subsets positive/disconnected and negative/disconnected. We then evaluate the performance of each model on those subsets. The tasks of signed link prediction between positive and zero samples are denoted as $p@z$ while the negative against zero is $n@z$. Furthermore, as the Signed Relational Latent Distance Model formulation facilitates the inference of a polytope describing the distinct aspects of networks, we visualize the latent space across various dimensions for all of the corresponding networks. To facilitate visualizations we use Principal Component Analysis (PCA), and project the space based on the first two principal components of the final embedding matrix. In addition, we provide circular plots where each archetype of the polytope is mapped to a circle every $\text{rad}_d = \frac{2\pi}{D}$ radians, with D being the number of archetypes. Polytope visualizations for multiple latent dimensions can be found in Figure 6.3.

During the evaluation, we focused on certain scoring metrics that are suitable for highly imbalanced data sets, specifically the AUC-ROC score and the AUC-PR score. We applied these scores to assess two particular tasks: link sign prediction and signed link prediction. In these tasks, we found that the Signed Hybrid-Membership Latent Distance Model performed competitively when measured against all baseline models. Specifically, in most cases, our framework outperformed significantly most baselines or defined on-par performance against the most competitive ones. Surprisingly, when

compared to the Skellam Latent Distance Model and the Signed Relational Latent Distance Model which define a higher model capacity our framework defined on-par performance. Controlling the volume of the simplex in the Signed Hybrid-Membership Latent Distance Model showed a small decrease in the predictive performance when the volume was decreased significantly. The type of memberships and participation across different archetypes became hard assignments when the volume again decreased significantly, showcasing a unique archetypal selection for the node reconstruction. Experiments on the two bipartite networks show that the model uncovered successful patterns modeling polarization of voting records both in the U.S. Congress and Senate, as seen by Figure 6.3. (For more details and the full experiment results please visit the full paper as provided in the Appendix in Section 9)

6.3 Conclusion

The Signed Hybrid-Membership Latent Distance Model allows for the archetypal characterization of signed networks even in the case where pure nodes are not present. Easily interpretable visualizations of signed networks are achieved by drawing the inferred latent space which in addition can provide more specialized interpretations as smaller volumes lead to node reconstructions from a unique archetype. Importantly, uniqueness guarantees allow for the robust interpretation of the inferred solution. Constraining the polytope to the D-simplex did not hamper the predictive performance but allowed for control over the type of memberships to the different archetypes.



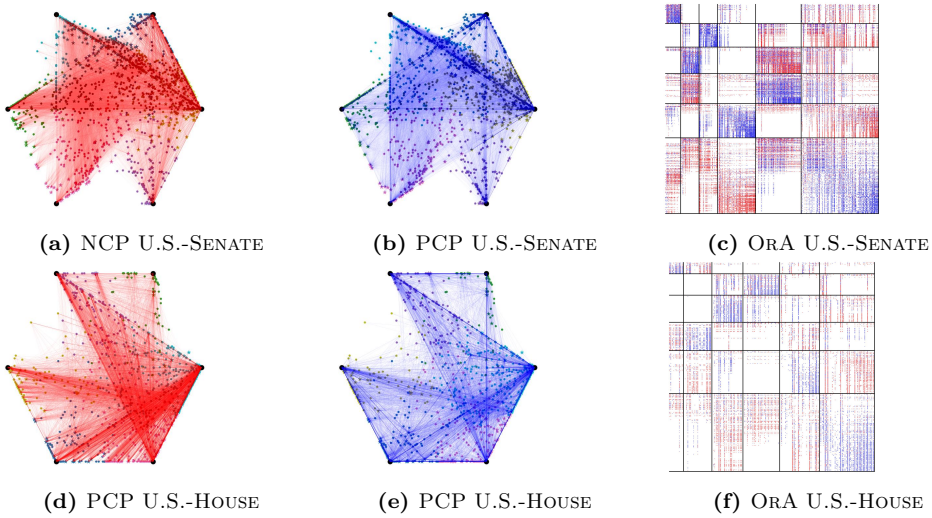
(a) PCA ($D = 8$) **(b)** PCA ($D = 16$) **(c)** PCA ($D = 32$) **(d)** PCA ($D = 64$)

(e) NCP ($D = 8$) **(f)** NCP ($D = 16$) **(g)** NCP ($D = 32$) **(h)** NCP ($D = 64$)

(i) PCP ($D = 8$) **(j)** PCP ($D = 16$) **(k)** PCP ($D = 32$) **(l)** PCP ($D = 64$)

(m) ORA ($D = 8$) **(n)** ORA ($D = 16$) **(o)** ORA ($D = 32$) **(p)** ORA ($D = 64$)

Figure 6.2. sHM-LDM(p=2): *Twitter* Network [153]—Inferred simplex visualizations and ordered adjacency matrices for various dimensions D and with simplex side lengths δ ensuring identifiability. The first row shows the latent space projection to the first two Principal Components—The second row provides a Negative Circular Plot (NCP) with red lines showcasing negative links between nodes—The third row shows a Positive Circular Plot (PCP) with the blue lines denoting positive links between node pairs—The fourth and final row shows the Ordered Adjacency (ORA) matrices sorted based on the memberships $\mathbf{w}_i$, in terms of maximum simplex corner responsibility, and internally according to the magnitude of the corresponding corner assignment for their reconstruction [100].



(a) NCP U.S.-Senate **(b)** PCP U.S.-Senate **(c)** OrA U.S.-Senate

(d) PCP U.S.-House **(e)** PCP U.S.-House **(f)** OrA U.S.-House

Figure 6.3. sHM-LDM(p=2): Inferred simplex visualizations and ordered adjacency matrices for a $D = 6$ dimensional simplex with side lengths δ ensuring identifiability. The first column provides a Negative Circular Plot (NCP) with red lines showcasing negative links between nodes—The second column shows a Positive Circular Plot (PCP) with the blue lines denoting positive links between node pairs—The third and final column shows the Ordered Adjacency (OrA) matrices ordered based on the memberships, in terms of maximum simplex corner responsibility, and internally according to the magnitude of the corresponding corner assignment for their reconstruction. Top row: U.S.-House [154]. Bottom row U.S.-Senate [154] [100].

Part IV

Graph Representation Learning of Single-Event Temporal Networks

CHAPTER 7

Time to Cite: Modeling Citation Networks using the Dynamic Impact Single-Event Embedding Model

A major focus has been given to the understanding of SciSci through the lens of complex network analysis, studying the structural properties and dynamics, of naturally occurring graph data describing SciSci. These include collaboration networks describing how scholars cooperate to advance various scientific fields. In particular, pioneering works [155–157] have analyzed multiple network statistics such as degree distribution, clustering coefficient, and average shortest paths. Furthermore, citation networks define an additional prominent case where graph structure data describe SciSci. Citation networks, essentially describe the directed relationships of papers (nodes) with an edge occurring between a dyad if paper A cites paper B, e.g. A→B. Multiple efforts towards Graph Representation Learning of citation networks have been made, although treating such networks as static in time. Notably, citation networks are dynamic. Whereas dynamic modeling approaches can uncover structures obscured when aggregating networks across time to form static networks, the dynamic modeling approaches are in general based on the assumption that multiple links occur between the dyads in time. Therefore no optimal likelihood formulation has been explored for such networks defined as single-event networks (SENs). Furthermore, lots of attention has been given in SciSci to the temporal impact characterization of papers in terms of their citation dynamics. Importantly, most of these studies relied on carefully designed heuristics that utilized classical machine learning methods based on various scholarly features, as well as, paper textual information. Such

features are used to quantify and predict a paper's impact included linear/logistic regression, k-nearest neighbors, support vector machines, random forests, and many more [158–163]. These studies focused primarily on carefully designing and including proper features to be used for the impact prediction task. Unfortunately, no method has successfully combined a Graph Representation Learning approach under an appropriate SEN likelihood while also accounting for impact characterization.

Consequently, we here focus on citation networks to alleviate such limitations. It is worth mentioning, that despite focusing only on citation graphs, our approach is eligible for the analysis of every network that falls under the SEN umbrella. We here turn to the Inhomogenous Poisson Point Process for which we constrain to the modeling of the maximum one event that may appear per dyad, yielding the Single-Event Poisson Process define for the first time a principled likelihood expression for single events networks. In order to define powerful ultra-low dimensional network embeddings we turn to the representation power of the directed network version LDM. Specifically, for every paper we define static embeddings distinguishing between source and target roles, i.e. we introduce a different position in the latent space for the roles of papers when citing or being cited. In addition, we define paper random effects that can be reparametrized to represent paper masses, again distinguishing between "being cited" and "citing" masses. For the "being cited" mass we introduce a temporal impact function that characterizes the incoming citation dynamics. eligible for impact quantification. The impact function is parameterized through appropriate probability density functions, including the log-normal, as well as, the truncated normal distributions.

7.1 Contributions

We presently extend the Latent Distance Model to account for single-event networks and to accurately characterize paper impact based on citation dynamics. We specify an Inhomogeneous Poisson Point Process for the analysis of SENs, defining the Single-Event Poisson Process which provides for the first time an appropriate likelihood for the SEN family of temporal networks. We hereby introduce the Dynamic Impact Single-Event Embedding Model (DISEE) which characterizes the scientific interactions in terms of a latent distance model in which forces (strength of the interaction) can be reparameterized to be proportional to the product of the masses of the interacting entities. To account for the time-varying impact, the mass of a contribution being used is time-dependent based on flexible parametric representations of scientific impact. The procedure overview is provided in Figure 7.1. Analytically our contributions are outlined as:

- We, for the first time, derive the single-event Poisson Process (SE-PP). As paper citation networks (and SENs in general) only include a single event we augment the Poisson Process likelihood to have support only for single events forming the single event Poisson Process.

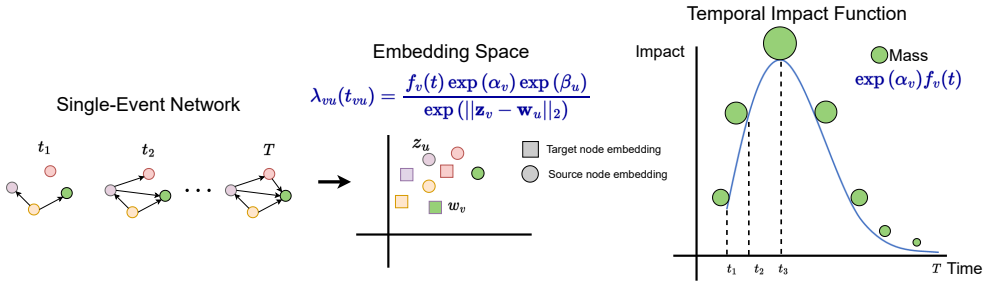


Figure 7.1. DISEE procedure overview. Given a Single-Event Network (SEN) as an input, the model defines an intensity function introducing two sets of static embeddings distinguishing between source $\mathbf{w}_u$ and target $\mathbf{z}_v$ node embeddings. Furthermore, each node is assigned its own random effect, distinguishing again the source β_u and target α_v roles. The random effects can be parameterized to represent source and target masses through the exponential function. Finally, for each target node of the network, the model further defines an impact function $f_v(t)$ yielding a temporal impact characterization of the nodes' incoming link dynamics which controls the nodes' mass in time as, $\exp\{(\alpha_v)\}f_v(t)$.

- We propose the Dynamic Impact Single-Event Embedding Model based on the SE-PP for SENs. We characterize the rate of interaction within a latent distance model such that citations are generated relative to the degree to which a paper cites and a paper is being cited at a given time point interpreted as masses of the citing and cited papers, respectively, augmented by their distance in latent space.

- We demonstrate how the Dynamic Impact Single-Event Embedding Model reconciles conventional impact modeling with latent distance embedding procedures. Specifically, we show how the model enables accurate dynamic characterization of citation impact similar to conventional paper impact modeling procedures while at the same time providing low-dimensional embeddings accounting for the structure of citation networks. We highlight this reconciliation on three real networks covering three distinct fields of science.

7.2 Experimental design, results, and key findings

We evaluate how successfully DISEE reconciles traditional impact quantification approaches with latent distance modeling. Specifically, we test the proposal the proposed approach's effectiveness in the link prediction task by comparing it to the classical LDM which is not time-aware and does not quantify temporal impact. We also consider multiple model ablations that are either able to characterize a node's impact or to account for GRL, i.e. define node embeddings, but not both. For the

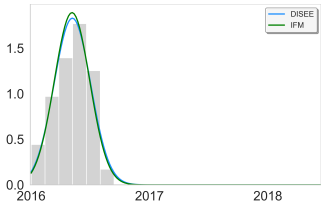
task of link prediction, we remove 20% of network links and we sample an equal amount of non-edges as negative samples and construct the test set. Notably, these negative samples are sampled in a time-aware manner, meaning that we consider only pairs that are possibly to exist as missing links in the network (i.e. we do not consider node pairs where missing citations refer to papers citing future papers, as the target paper did not exist the time when the source paper was published). The link removal is designed in such a way that the residual network stays connected. Analytically, for each network, we do not consider removing links that make up the minimum spanning tree of the graph. For the evaluation, we consider both the Receiver Operator Characteristic and Precision-Recall Area Under Curve scores, as these are metrics not sensitive to the class imbalance between links and non-links. We then continue by evaluating the quality of impact expression of DISEE by visually presenting the inferred impact functions and comparing them against an Impact Function Model (IFM) which fits an impact function directly on the citation pattern of each paper. Finally, we visualize the model's learned temporal space representing the target papers, accounting for their temporal impact in terms of their mass at a specific time point, and characterizing the different papers' lifespans.

For the link prediction experiments, the best performance is achieved by model specifications that define an embedding space, i.e. the DISEE and LDM models while the rest of the model ablations defined significantly lower performance. Comparing the two distribution choices for the impact function (Truncated Normal and Log Normal) we observed very similar link prediction scores. We continued by addressing the quality of paper impact characterization based on a target paper's incoming citation dynamics. In such a direction, we further compared the inferred impact functions of the *DISEE* and IFM, under the Truncated normal and Log Normal distributions, against the true impact dynamics for each one of the corresponding papers. For the Truncated case (Figure 7.2), we observe that *DISEE* and IFM provide very similar (and in some cases identical) impact functions that capture the underlying citation patterns. In the case of the Log-Normal distribution (Figure 7.3), we witness an agreement between DISEE and IFM models when the paper lifespan does not exceed the 2 years. For larger lifespans DISEE defines a larger standard deviation than the IFM returning much heavier tails. Both models when compared to the true citation histogram provide much heavier tails when the paper lifespan exceeds the 2-year threshold. The Log-Normal distribution is not invariant to the scale of the x-axis (contrary to the Truncated normal which is scale-invariant) and this can be potentially a reason for observing this kind of behavior, meaning that the choice of the time resolution is not optimal (this is to be further investigated). Nevertheless, the Truncated normal distribution seems to very accurately represent the true citation dynamics, defining correct distribution tails, but in some cases, the Log-Normal heavier tails may be more appropriate for future impact predictions (as papers stay "alive" longer). Finally, we provide embedding space visualizations of the target (cited) papers, accounting for their temporal impact in terms of their mass at a specific time point, showcasing the evolution of the embedding space for the domain of *Machine Learning*. These visualizations showed that as the years progress,

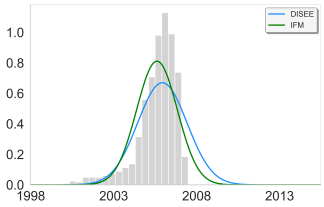
paper masses reach much larger magnitudes than in the earlier years, defining higher research significance, and accumulating higher citation numbers and impact which can be explained by the increase in published *Machine Learning* works. Embedding space visualizations are provided in Figures 7.4 and 7.5. (For more details and the full experiment results please visit the full paper as provided in the Appendix in Section 9)

7.3 Conclusion

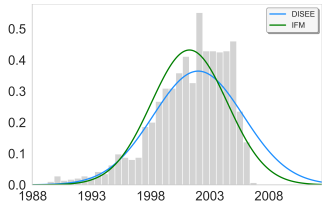
We have proposed the Dynamic Impact Single-Event Embedding Model (DISEE), a reconciliation between traditional impact quantification approaches with a Latent Distance Model (LDM). We have focused on Single-Event Networks (SENs), and more specifically in citation networks, where we for the first time derived Single-Event Poisson Process. Such a process defines an appropriate likelihood allowing for a principled analysis of single-events networks. In order to define powerful ultra-low dimensional network embeddings we turn to the representation power of the directed network version of the LDM. Specifically, for every paper we define static embeddings distinguishing between source and target roles, i.e. we introduced a different position in the latent space for the roles of papers when citing or being cited. In addition, we defined paper random effects that can be reparametrized to represent paper masses, again distinguishing between ”being cited” and ”citing” masses. For the ”being cited” mass we introduced a temporal impact function that characterized the incoming citation dynamics. eligible for impact quantification. The impact function is parameterized through appropriate probability density functions, including the log-normal, as well as, the truncated normal distributions. Through extensive experiments, we showed that the DISEE had the same link prediction performance as the powerful LDM. Furthermore, we showed that the temporal impact characterization was validated by an Impact Function Model IFM. These results, showcase that the DISEE successfully reconciles powerful embedding approaches with citation dynamics impact characterization. Finally, visualizations of the embedding space for target papers provided accurate representations that described the birth and death of papers following their impact lifespans as years pass and science moves forward.



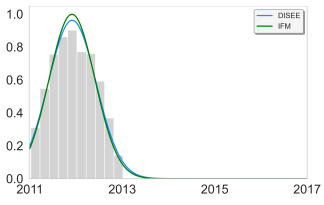
(a) Deep Residual Learning for Image Recognition [164].



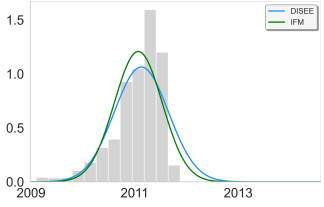
(b) Gradient-based learning applied to document recognition [165].



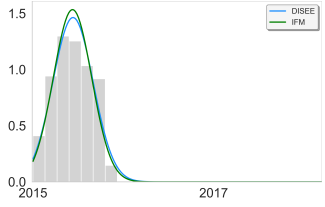
(c) Structural equation modeling in practice [166].



(d) LIBSVM: A library for support vector machines [167].

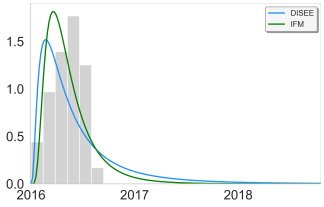


(e) ImageNet: A large-scale hierarchical image database [**?**].

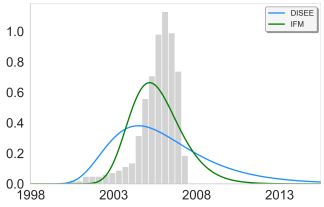


(f) Going deeper with convolutions [168].

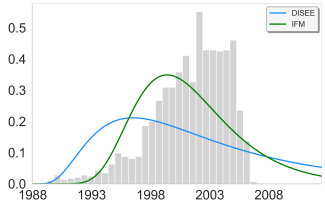
Figure 7.2. *Machine Learning*: DISEE Truncated and IFM Trunctated models inferred impact function visualizations compared to the true citation histogram, for six popular *Machine Learning* papers with different citation dynamics.



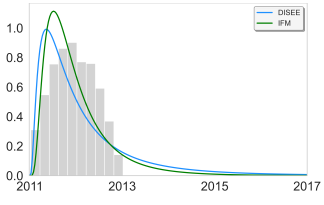
(a) Deep Residual Learning for Image Recognition.



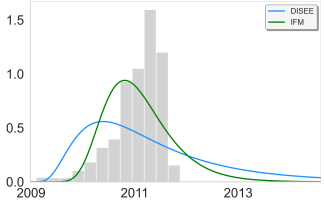
(b) Gradient-based learning applied to document recognition.



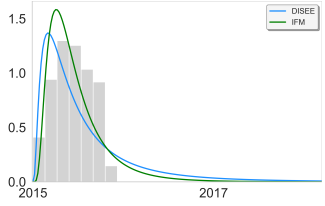
(c) Structural equation modeling in practice.



(d) LIBSVM: A library for support vector machines.



(e) ImageNet: A large-scale hierarchical image database.



(f) Going deeper with convolutions.

Figure 7.3. *Machine Learning*: DISEE Log Normal and IFM Log Normal models inferred impact function visualizations compared to the true citation histogram, for six popular *Machine Learning* papers with different citation dynamics.

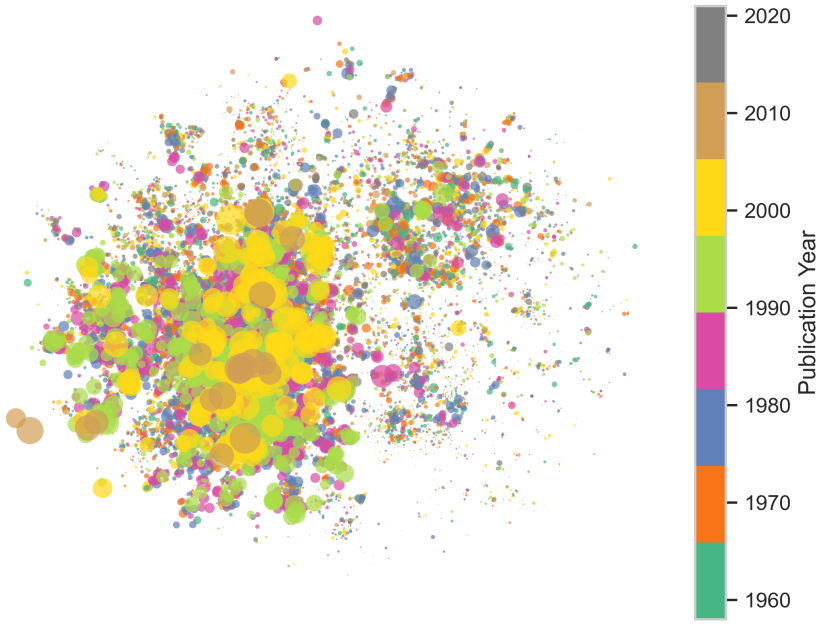


Figure 7.4. *Machine Learning*: DISEE TRUNCATED embedding space visualization for all target papers published before the year 2023. Node sizes are based on each paper's current mass, $f_i(t) * \exp\{(\alpha_i)\}$, and thus papers with zero mass are not visible denoting the end of their scientific relevance or "lifespan". Nodes are color-coded based on their publication year.

(a) 1988 **(b)** 1998 **(c)** 2008 **(d)** 2018

Figure 7.5. *Machine Learning*: DISEE TRUNCATED embedding space evolution throughout the years. Node sizes are based on each paper's mass, $f_i(t) * \exp\{(\alpha_i)\}$, showcasing how papers reach the end of their scientific relevance or "lifespan" by disappearing from the embedding space as time progresses. Nodes are color-coded based on their publication year.

Part V

Discussion and conclusion

CHAPTER 8

Discussion

Our work aimed to create novel Graph Representation Learning approaches, and most of all we tried to define what constitutes a fine embedding approach for accurate graph representation. We argue the characteristics of a fine embedding approach should 1) Be interpretable by human perception, similar nodes should be positioned in close proximity in the latent space, i.e. node similarity on the network should be translated into similarity in the latent space (one of the main goals, and intuition behind GRL). 2) Provide insights over the intrinsic structures existing in the network, facilitating interpretation and visualization in a hierarchical/multi-resolutional manner or even extracting pure network nodes and extreme profiles, characterizing network polarization. 3) Visualizations should not depend on heuristic dimensionality reduction approaches but provide accurate low-dimensional representations with maximum $D = 3$. 4) Return high performance in downstream tasks such as link prediction/network reconstruction/node classification and community detection. 5) Scale the analysis to massive and large-scale networks, as billion-node graphs become more and more common in real scenarios.

We demonstrated how the proposed frameworks provide such fine network representations since (1) They operate under the Euclidean distance metric, conveying homophily and transitivity properties, and thus providing an intuitive human perception of both first and high-order node similarity. (2) Naturally characterizing network intrinsic structures via the use of multi-scale hierarchical block structures, or constrained-to-polytopes latent spaces, providing hierarchical community identification, hybrid community memberships, and uncovering of extreme node profiles. 3) All of the frameworks showed very competitive performance under ultra-low dimension of $D = 2, 3$, providing direct network visualizations but sufficient capacity to enable accurate interpretation of network images. 4) High performance was achieved by all of the proposed methods in multiple downstream tasks, outperforming most of the state-of-the-art baselines while also enabling generative processes contrary to most of the competing methods. 5) Our methods accounted for the computational costs of modern large-scale networks defining accurate linearithmic approximations of the network likelihood, unbiased random sampling procedures, and case-control inferences.

The first phase of this thesis focused on the Graph Representation Learning of positive integer weighted graphs. Most of all we tried to define what constitutes a fine embedding approach for accurate graph representation. Analytically, we first

developed the Hierarchical Block Distance Model (HBDM), a scalable reconciliation of latent distance models and their ability to account for homophily and transitivity with hierarchical representations of network structures. We demonstrated how the proposed HBDM provides favorable network representations by (1) Operating with a Euclidean distance metric providing an intuitive human perception of node similarity. (2) Naturally representing multiscale hierarchical structure based on its block structure and carefully designed clustering procedure optimized in terms of Euclidean distances. (3) Directly and consistently operating in $D = 2, 3$ with high performance. (4) Performing well on all considered downstream tasks highlighting its ability to account for the underlying network structure. Importantly, the inferred hierarchical structure admits community discovery at multiple scales as highlighted by the inferred dendrograms and ordered adjacency matrices, and naturally extends to the characterization of communities of bipartite networks. Our discoveries highlight the existence and importance of hierarchical multi-scale structures in complex networks. The across hierarchy re-ordered adjacency matrices given by HBDM, manifest subcommunities inside of what already appears as a strongly connected community. This points to how delicate the task of defining communities is and the importance of accounting for communities at multiple scales, as enabled by the HBDM. Importantly, these results generalize for bipartite networks where multi-scale geometric representations, joint hierarchical structures, and community discovery are arduous tasks. In conclusion, we proposed the Hierarchical Block Distance Model, a scalable reconciliation of network embeddings using the latent distance model (LDM) and hierarchical characterizations of structure at multiple scales via a novel clustering framework. Notably, the model mimics the behavior of the LDM where the use of homophily and transitivity is most important while scaling in complexity by $\mathcal{O}(DN \log N)$. We analyzed thirteen networks from moderate sizes to large-scale with the HBDM having favorable performance when compared to existing scalable embedding procedures. In particular, we observed that the HBDM well predicts links and node classes utilizing a very low embedding dimension of $D = 2$ providing accurate network visualizations and characterization of structure at multiple scales. Our results demonstrate that favorable performance can be achieved using ultra-low (i.e. $D = 2$) embedding dimensions and a scalable hierarchical representation that accounts for homophily and transitivity.

In the same direction, we have proposed the Hybrid-Membership Latent Distance Model (HM-LDM) that reconciles network embedding and latent community detection. The approach utilizes both the normal and squared Euclidean distance model where the latter integrated the non-negativity-constrained Eigenmodel with the Latent Distance Model. We demonstrated that the model could be constrained to the simplex without losing expressive power. The reduced simplex provides unique representations, ultimately resulting in the hard clustering of nodes to communities when the simplex is sufficiently shrunk. Notably, the proposed HM-LDM combines network homophily and transitivity properties with latent community detection enabling explicit control of soft and hard assignment through the volume of the induced simplex. We observed favorable link prediction performance in regimes in which the HM-

LDM provides unique representations while enabling the ordering of the adjacency matrix in terms of prominent latent communities. Finally, we showed the ability of the model to extract valid community structures across multiple networks and showcased how the analysis extends to bipartite networks. Future work should compare the performance of HM-LDM against classical non-embedding methods such as the Degree Corrected Stochastic Block Model (DC-SBM) [118] or the Mixed Membership Stochastic Block Model (MM-SBM) [169]. Such a comparison is of particular interest since DC-SBM accounts for degree heterogeneity while MM-SBM for soft assignments, two important properties of HM-LDM.

In the second phase of the thesis, we focused on the analysis of signed integer-weighted networks. In that direction, we proposed the Skellam Latent Distance Model (SLDM) and Signed Latent Relational Distance model (SLIM) to provide easily interpretable network visualization with favorable performance in the link prediction tasks for weighted signed networks. In particular, endowing the model with a space-constrained to polytopes (forming the Signed relational Latent dIstance Model(SLIM)) enabled us to characterize distinct aspects in terms of extreme positions in the social networks akin to conventional archetypal analysis but for graph-structured data. The Skellam distribution is considerably beneficial in modeling signed networks, whereas the relational extension of AA can be applied for other likelihood specifications, such as LDMs in general. This work thereby provides a foundation for using likelihoods accommodating weighted signed networks and representations akin to AA in general for analyzing networks.

Later in the second phase, we presented the signed Hybrid-Membership Latent Distance Model (sHM-LDM) reconciling Graph Representation Learning and latent community detection in singed networks. Specifically, we extended a hybrid membership model to account for signed networks and showed that a minimum volume approach could uncover distinct profiles in social networks while ensuring model identifiability. The presented framework was formulated to include an Euclidean as well as a squared Euclidean norm. For the latter, a direct relationship to an Eigenmodel was shown. Furthermore, by controlling the volume of the simplex by the magnitude of δ, a sufficiently reduced simplex leads to unique representations. Notably, the generalization to signed networks facilitated the extraction of distinct network profiles representing positive interactions and animosity. In regimes where the sHM-LDM provide unique representations, we observed favorable link prediction performance and the ability to order the adjacency matrix based on prominent latent communities and distinct profiles. Importantly, the extended sHM-LDM merges homophily and heterophily properties to account for positive and negative ties as present in signed networks, enabling explicit control of soft and hard assignment to extreme node profiles, through the volume of the induced simplex.

In the third phase, we focused on SEN networks and more specifically on the analysis of citation networks. We proposed a novel likelihood function for the characterization of such single-event networks. Using this likelihood, we defined the Dynamic Impact Single-Event Embedding Model (DISEE) characterizing scientific interactions and impact, in terms of a latent distance model in which forces were reparameterized

to be proportional to the product of the masses of the interacting entities. Such a model successfully reconciled static latent distance network embedding approaches with classical dynamic impact assessments of citation networks. Extensive experiments in three real citation networks, showcased DISEE as a powerful link predictor, able to successfully describe papers' impact and relevance lifespans while visualization of the inferred embedding space provided new insights on how different domains of science evolve through time.

Our finding of ultra-low dimensional accurate characterizations of network structures supports the findings in [170] in which a logistic PCA model was found to enable exact low-dimensional recovery of multiple real-world networks. Whereas the work of [170] focuses on exact network reconstruction we find that generalizable patterns can be well extracted in ultra-low dimensional representations with performance saturating after just $D = 8$ dimensions for all networks considered. Whereas [170] found that their low-dimensional space did not perform well in classification tasks we observed strong node classification performance by the low-dimensional representations provided by our frameworks. Importantly, in node classification and the HBDM, we observed better performance using KNN as opposed to simple linear classification based on logistic/multinomial regression typically used for node classification. This highlights that whereas most GRL works use linear classifiers there is no guarantee that the embedding space will be linearly separable and performance should therefore be compared to non-linear classifiers as they may provide more favorable performance as observed in this study.

Recent pioneering works [171,172] have drawn significant attention of the research community by questioning the conventional embedding space preference, as also reviewed for the LSM family [71]. It is well known that many real-world networks show power-law degree distribution, or they can consist of latent hierarchical inner structures. Therefore, Euclidean space might not always be appropriate to represent such complex network architectures. It might also require higher-dimensional spaces to show comparable performance in the GRL tasks. The works of [171, 172] demonstrated that hyperbolic spaces, such as the Poincare disk model, can provide substantial benefits over the Euclidean space. The presented models, naturally extend to other distance measures and future studies should explore how they can be extended to hierarchical representations and polytopes-defined spaces beyond Euclidean geometry.

Covariate information plays an important role in the outstanding performance of GRL methods, especially GNNs. In the current LSM literature, side information is accounted for by extra regressors in the logit/log link functions expressing the likelihood of a dyad being connected. Using the Mahalanobis distance imposing a block-diagonal covariance matrix, the proposed frameworks can naturally incorporate covariate information directly into the latent space and notably construct multi-scale structures and polytope representations via the enriched and concatenated embedding of the latent variables and the covariate information. In more detail, we can define a new embedding matrix $\bar{\boldsymbol{Z}}$ as the concatenation over the latent variables and the covariate information for node i as: $\bar{\boldsymbol{z}}_i = [\boldsymbol{z}_i; \boldsymbol{x}_i]$ and a Mahalanobis correlation matrix

as: $\boldsymbol{S} = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0}^T & \boldsymbol{J} \end{bmatrix} \in \mathbb{R}^{(D+R)\times(D+R)}$, where $I \in \mathbb{R}^{D\times D}$ the identity matrix, the zero matrix $\boldsymbol{0} \in \mathbb{R}^{D\times R}$ and the covariate coefficient matrix $\boldsymbol{J} \in \mathbb{R}^{R\times R}$. In this setting, HBDM is able to construct a covariate information-aware multi-scale latent space by the use of the Mahalanobis distance $d_{ij} = \sqrt{(\bar{\boldsymbol{z}}_i - \bar{\boldsymbol{z}}_j)^T \boldsymbol{S}^{-1} (\bar{\boldsymbol{z}}_i - \bar{\boldsymbol{z}}_j)}$. Our analysis presently did not explore side information and this is also why we did not include comparisons to prominent GNN-based approaches as these procedures do not provide favorable performance when only learning from the graph structure itself. As such, we observed (not shown) poor performance of GraphSage [38] when only having access to the graph structure in the present setup. Our presented methods, operate on static networks and thus are not naturally inductive models. Nevertheless, potential new emerging nodes can be projected into the inferred latent space by fixing the embeddings of nodes present in the training set while optimizing the new nodes for their locations in the learned latent space. We leave a comparison of such a strategy against naturally inductive models such as GNNs for future work.

Our frameworks, use the LDM and thus are good at characterizing transitivity and homophily at a node and cluster level, whereas the random effects enable accounting for degree heterogeneity. Notably, our methods suffer from the limitations of the LDM and are thus unable to model stochastic equivalence. Future work should therefore investigate hierarchical structures and polytope representations imposed on more flexible GRL procedures enabling stochastic equivalence and contrast the performance when accounting for stochastic equivalence to the existing methods based on the SBM which as a latent class model is known to express stochastic equivalence [93,173–177]. In addition, the optimization for our frameworks is a highly non-convex problem and thus relies on the quality of initialization in terms of convergence speed. In this regard, we use a deterministic initialization based on the normalized Laplacian. In addition, for the signed network models we observed that a maximum likelihood estimation of the model parameters became unstable when the network contained some nodes having only negative interactions. This is a direct consequence of the presence of the distance term $(\exp(+||\cdot||_2))$ for negative interactions, which can lead to overflow during inference. Nevertheless, we adopted a MAP estimation that was found to be stable across all networks. For real signed networks, the generative model created an "excess" of negative links increasing the overall network sparsity. For that, a modified *SLIM* excluding the regularization over the model parameters was introduced which achieved correct network sparsity (as shown in the main paper). Assuming priors over the model parameters created a bias over the generated network when compared to the ground truth network statistics.

CHAPTER 9

Conclusion

In recent years, there has been a surge in the complexity and volume of data represented as graphs. In this context, we have presented innovative representation learning models, based on the Latent Distance Model formulation, specifically tailored for the examination of networks that involve both signed and unsigned integer weights, as well as, single-event networks. We have successfully presented multiple frameworks able to learn informative node representations, expressing homophily and transitivity properties in unsigned networks while for the case of signed networks, models were generalized to convey the balance theory. The Hierarchical Block Distance Model facilitated the extraction of hierarchical structures present in complex networks while the Hybrid Membership Distance Model accounted for community discovery, explicitly controlling both hard and soft community assignments. Furthermore, the family of Latent Distance Models was extended to the analysis of signed networks via the Skellam Latent Distance Model which was proved to be a powerful link predictor. Constraining the latent space to a polytope yielded the Signed Relational Latent Distance model generalizing Archetypal Analysis to relational data extracting distinct profiles of networks and characterizing network polarization. When the polytope was constrained to the D-simplex we obtained the signed Hybrid-Membership Latent Distance Model which a continuously decreasing simplex volume, defined a Minimum Volume approach for Archetypal Analysis yielding also extreme profile identification. Importantly, all proposed frameworks defined scalable optimization approaches via the accurate linearithmic hierarchical approximation of the likelihood (HBDM), unbiased random sampling procedures (HM-LDM, sHM-LDM, SLDM, SLIM), and case-control inferences (DISEE). Our frameworks facilitated informative network visualizations including network hierarchical organization of the adjacency matrix (HBDM), soft and hard community extraction (HM-LDM), informative polytope visualizations for signed networks (sHM-LDM, SLIM), and impact characterization and latent space visualizations of single-event networks (DISEE). Importantly, such valuable visualization analyses were extended to bipartite networks where such a generalization is not trivial. For all of our proposed models, we included extensive experimental evaluations to demonstrate that the proposed approaches generally surpass widely adapted baseline methods in node classification, link prediction, and network reconstruction tasks. Such results were highlighted especially for the ultra-low dimensions of $D = 2, 3$ where very few of the competing methods were found to be competitive. Importantly, the proposed frameworks were validated in multiple set-

tings and downstream tasks and were found to have the most consistent performance across tasks (in no task their performance was significantly lower than competing baselines). This helps us to characterize embedding approaches relying on the Euclidean metric as the best choice when it comes to defining low-dimensional embeddings that are required to perform multiple tasks. Last but not least, we have successfully shed light on a missing part in the GRL literature which is to extensively position and benchmark the performance of Latent Distance Models for Graph Representation Learning against state-of-the-art baselines, showcasing their superior performance in multiple settings.

Bibliography

[1] M. E. J. Newman, "The structure and function of complex networks," *SIAM Review*, vol. 45, no. 2, pp. 167–256, 2003.

[2] D. Liben-Nowell and J. Kleinberg, "The link prediction problem for social networks," in *CIKM*, p. 556–559, 2003.

[3] L. Getoor and B. Taskar, *Introduction to Statistical Relational Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2007.

[4] A. Grover and J. Leskovec, "Node2Vec: Scalable feature learning for networks," in *KDD*, pp. 855–864, 2016.

[5] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, no. 3, pp. 75–174, 2010.

[6] B. S. Khan and M. A. Niazi, "Network community detection: A review and visual survey," *CoRR*, vol. abs/1708.00977, 2017.

[7] A. Vespignani, "Twenty years of network science," *Nature*, vol. 558, pp. 528 – 529, 2018.

[8] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, pp. 440–442, 1998.

[9] A.-L. Baraba

si and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, pp. 509–512, oct 1999.

[10] P. L. Erdos and A. Rényi, "On random graphs. i.," *Publicationes Mathematicae Debrecen*, 2022.

[11] J. E. Cohen, "Infectious Diseases of Humans: Dynamics and Control," *JAMA*, vol. 268, pp. 3381–3381, 12 1992.

[12] S. V. Buldyrev, R. Parshani, G. Paul, H. E. Stanley, and S. Havlin, "Catastrophic cascade of failures in interdependent networks," *Nature*, vol. 464, pp. 1025–1028, Apr. 2010.

[13] M. Kitsak, L. K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. E. Stanley, and H. A. Makse, "Identification of influential spreaders in complex networks," *Nature Physics*, vol. 6, pp. 888–893, Aug. 2010.

[14] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences*, vol. 99, pp. 7821–7826, June 2002.

[15] A.-L. Barabási and M. Pósfai, *Network science*. Cambridge University Press, 2016.

[16] L. Freeman, "A set of measures of centrality based on betweenness," *Sociometry*, vol. 40, pp. 35–41, 03 1977.

[17] M. E. J. Newman, *Mathematics of Networks*, pp. 1–8. London: Palgrave Macmillan UK, 2016.

[18] L. Lü and T. Zhou, "Link prediction in complex networks: A survey," *Physica A: Statistical Mechanics and its Applications*, vol. 390, pp. 1150–1170, mar 2011.

[19] L. Katz, "A new status index derived from sociometric analysis," *Psychometrika*, vol. 18, pp. 39–43, Mar. 1953.

[20] E. A. Leicht, P. Holme, and M. E. J. Newman, "Vertex similarity in networks," *Physical Review E*, vol. 73, feb 2006.

[21] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking : Bringing order to the web," in *The Web Conference*, 1999.

[22] B. Perozzi, R. Al-Rfou, and S. Skiena, "DeepWalk: Online learning of social representations," *CoRR*, vol. abs/1403.6652, 2014.

[23] Jianbo Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[24] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, NIPS'01, (Cambridge, MA, USA), p. 849–856, MIT Press, 2001.

[25] J. B. Tenenbaum, V. d. Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[26] *Multidimensional Scaling*. Sage Publications, 1978.

[27] J. Wang, *Laplacian Eigenmaps*, pp. 235–247. Springer Berlin Heidelberg, 2012.

[28] F. R. K. Chung, *Spectral Graph Theory*. American Mathematical Society, 1997.

[29] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," *Advances in Neural Information Processing System*, vol. 14, 04 2002.

[30] X. He and P. Niyogi, "Locality preserving projections," in *Advances in Neural Information Processing Systems* (S. Thrun, L. Saul, and B. Schölkopf, eds.), vol. 16, MIT Press, 2004.

[31] B. Shaw and T. Jebara, "Structure preserving embedding," in *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, p. 937–944, Association for Computing Machinery, 2009.

[32] Y. Yang, F. Nie, S. Xiang, Y. Zhuang, and W. Wang, "Local and global regressive mapping for manifold learning with out-of-sample extrapolation.," vol. 1, 01 2010.

[33] A. Ahmed, N. Shervashidze, S. Narayanamurthy, V. Josifovski, and A. J. Smola, "Distributed large-scale natural graph factorization," in *Proceedings of the 22nd International World Wide Web Conference (WWW 2013)*, 2013.

[34] W. L. Hamilton, "Graph representation learning," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 14, no. 3, pp. 1–159.

[35] D. Zhang, J. Yin, X. Zhu, and C. Zhang, "Network representation learning: A survey," *IEEE Transactions on Big Data*, vol. 6, no. 1, pp. 3–28, 2020.

[36] W. L. Hamilton, R. Ying, and J. Leskovec, "Representation learning on graphs: Methods and applications," *IEEE Data Eng. Bull.*, vol. 40, no. 3, pp. 52–74, 2017.

[37] J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, and Q. Mei, "LINE: Large-scale information network embedding," in *WWW*, pp. 1067–1077, 2015.

[38] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," *Advances in neural information processing systems*, vol. 30, 2017.

[39] M. Zhang and Y. Chen, "Link prediction based on graph neural networks," 2018.

[40] N. Shervashidze, P. Schweitzer, E. J. van Leeuwen, K. Mehlhorn, and K. M. Borgwardt, "Weisfeiler-lehman graph kernels," *Journal of Machine Learning Research*, vol. 12, no. 77, pp. 2539–2561, 2011.

[41] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE Transactions on Neural Networks*, vol. 20, pp. 61–80, 2009.

[42] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2017.

[43] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," 2018.

[44] Y. Zheng, S. Chen, X. Zhang, X. Zhang, X. Yang, and D. Wang, "Heterogeneous-temporal graph convolutional networks: Make the community detection much better," 2020.

[45] J. Chen, T. Ma, and C. Xiao, "Fastgcn: Fast learning with graph convolutional networks via importance sampling," 2018.

[46] J. Qiu, Y. Dong, H. Ma, J. Li, K. Wang, and J. Tang, "Network embedding as matrix factorization: Unifying DeepWalk, LINE, PTE, and Node2Vec," in *WSDM*, 2018.

[47] J. Qiu, Y. Dong, H. Ma, J. Li, C. Wang, K. Wang, and J. Tang, "NetSMF: Large-scale network embedding as sparse matrix factorization," in *WWW*, pp. 1509–1520, 2019.

[48] S. Cao, W. Lu, and Q. Xu, "GraRep: Learning graph representations with global structural information," in *CIKM*, pp. 891–900, 2015.

[49] M. Ou, P. Cui, J. Pei, Z. Zhang, and W. Zhu, "Asymmetric transitivity preserving graph embedding," in *KDD*, pp. 1105–1114, 2016.

[50] J. Zhang, Y. Dong, Y. Wang, J. Tang, and M. Ding, "Prone: Fast and scalable network representation learning," in *IJCAI*, pp. 4278–4284, 7 2019.

[51] A. K. Bhowmick, K. Meneni, M. Danisch, J.-L. Guillaume, and B. Mitra, "LouvainNE: Hierarchical louvain method for high quality and scalable network embedding," in *WSDM*, pp. 43–51, 2020.

[52] H. Chen, B. Perozzi, Y. Hu, and S. Skiena, "HARP: hierarchical representation learning for networks," in *AAAI*, pp. 2127–2134, 2018.

[53] Z. Zhang, P. Cui, H. Li, X. Wang, and W. Zhu, "Billion-scale network embedding with iterative random projection," in *ICDM*, pp. 787–796, 2018.

[54] L. Lee, "Foundations of Statistical Natural Language Processing," *Computational Linguistics*, vol. 26, pp. 277–279, 06 2000.

[55] A. Çelikkanat and F. D. Malliaros, "Exponential family graph embeddings," in *AAAI*, pp. 3357–3364, 2020.

[56] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *NIPS*, pp. 3111–3119, 2013.

[57] D. Nguyen and F. D. Malliaros, "BiasedWalk: Biased sampling for representation learning on graphs," in *Big Data*, pp. 4045–4053, 2018.

[58] C. Li, S. Wang, D. Yang, Z. Li, Y. Yang, X. Zhang, and J. Zhou, "Ppne: Property preserving network embedding," in *Database Systems for Advanced Applications* (S. Candan, L. Chen, T. B. Pedersen, L. Chang, and W. Hua, eds.), (Cham), pp. 163–179, Springer International Publishing, 2017.

[59] J. Li, J. Zhu, and B. Zhang, "Discriminative deep random walk for network classification," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, (Berlin, Germany), pp. 1004–1013, Association for Computational Linguistics, Aug. 2016.

[60] S. Pan, J. Wu, X. Zhu, C. Zhang, and Y. Wang, "Tri-party deep network representation," in *International Joint Conference on Artificial Intelligence*, 2016.

[61] W. L. Hamilton, R. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *NIPS*, 2017.

[62] D. K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik, and R. P. Adams, "Convolutional networks on graphs for learning molecular fingerprints," in *Advances in Neural Information Processing Systems* (C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, eds.), vol. 28, Curran Associates, Inc., 2015.

[63] M. Henaff, J. Bruna, and Y. LeCun, "Deep convolutional networks on graph-structured data," *CoRR*, vol. abs/1506.05163, 2015.

[64] J. Bruna, W. Zaremba, A. Szlam, and Y. LeCun, "Spectral networks and locally connected networks on graphs," 2014.

[65] Matias, Catherine and Robin, Stéphane, "Modeling heterogeneity in random graphs through latent space models: a selective review*," *ESAIM: Proc.*, vol. 47, pp. 55–74, 2014.

[66] D. K. Sewell and Y. Chen, "Latent space models for dynamic networks," *Journal of the American Statistical Association*, vol. 110, no. 512, pp. 1646–1657, 2015.

[67] M. Salter-Townshend and T. H. McCormick, "Latent space models for multiview network data," *The Annals of Applied Statistics*, vol. 11, no. 3, pp. 1217 – 1244, 2017.

[68] L. Zhu, D. Guo, J. Yin, G. V. Steeg, and A. Galstyan, "Scalable temporal latent space inference for link prediction in dynamic social networks," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 10, pp. 2765–2777, 2016.

[69] A. Çelikkanat, N. Nakis, and M. Mørup, "Piecewise-velocity model for learning continuous-time dynamic node representations," in *Proceedings of the First Learning on Graphs Conference* (B. Rieck and R. Pascanu, eds.), vol. 198 of *Proceedings of Machine Learning Research*, pp. 36:1–36:21, PMLR, 09–12 Dec 2022.

[70] P. Sarkar and A. W. Moore, "Dynamic social network analysis using latent space models," in *NIPS*, pp. 1145–1152, 2006.

[71] A. L. Smith, D. M. Asta, and C. A. Calder, "The Geometry of Continuous Latent Space Models for Network Data," *Statistical Science*, vol. 34, no. 3, pp. 428 – 453, 2019.

[72] J. Sosa and L. Buitrago, "A review of latent space models for social networks," *CoRR*, vol. abs/2012.02307, 2020.

[73] B. Kim, K. Lee, L. Xue, and X. Niu, "A review of dynamic network models with latent variables," 2017.

[74] N. Nakis, A. Çelikkanat, and M. Mørup, "Hm-ldm: A hybrid-membership latent distance model," 2022.

[75] N. Nakis, A. Çelikkanat, L. Boucherie, C. Djurhuus, F. Burmester, D. M. Holmelund, M. Frolcová, and M. Mørup, "Characterizing polarization in social networks using the signed relational latent distance model," 2023.

[76] P. D. Hoff, A. E. Raftery, and M. S. Handcock, "Latent space approaches to social network analysis," *JASA*, vol. 97, no. 460, pp. 1090–1098, 2002.

[77] H. Louch, "Personal network integration: transitivity and homophily in strong-tie relations," *Social Networks*, vol. 22, no. 1, pp. 45–64, 2000.

[78] C. Zhang, Y. Bu, Y. Ding, and J. Xu, "Understanding scientific collaboration: Homophily, transitivity, and preferential attachment," *J. Assoc. Inf. Sci. Technol.*, vol. 69, p. 72–86, jan 2018.

[79] P. BLOCK and T. GRUND, "Multidimensional homophily in friendship networks," *Network Science*, vol. 2, no. 2, p. 189–212, 2014.

[80] *Exponential Random Graph Models for Social Networks: Theory, Methods, and Applications*. Structural Analysis in the Social Sciences, Cambridge University Press, 2012.

[81] P. N. Krivitsky, M. S. Handcock, A. E. Raftery, and P. D. Hoff, "Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models," *Social Networks*, vol. 31, no. 3, pp. 204 – 213, 2009.

[82] D. Cartwright and F. Harary, "Structural balance: a generalization of heider's theory.," *Psychological review*, vol. 63 5, pp. 277–93, 1956.

[83] D. Liben-Nowell and J. Kleinberg, "The link prediction problem for social networks," in *Proceedings of the Twelfth International Conference on Information and Knowledge Management*, CIKM '03, (New York, NY, USA), p. 556–559, Association for Computing Machinery, 2003.

[84] A. Sankar, Y. Wu, L. Gou, W. Zhang, and H. Yang, "Dynamic graph representation learning via self-attention networks," 2018.

[85] D. M. Dunlavy, T. G. Kolda, and E. Acar, "Temporal link prediction using matrix and tensor factorizations," *ACM Transactions on Knowledge Discovery from Data*, vol. 5, pp. 1–27, feb 2011.

[86] M. Nickel, V. Tresp, and H.-P. Kriegel, "A three-way model for collective learning on multi-relational data," in *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, (Madison, WI, USA), p. 809–816, Omnipress, 2011.

[87] L. Zhu, D. Guo, J. Yin, G. V. Steeg, and A. Galstyan, "Scalable temporal latent space inference for link prediction in dynamic social networks," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 10, pp. 2765–2777, 2016.

[88] C. Blundell, J. Beck, and K. A. Heller, "Modelling reciprocating relationships with hawkes processes," in *NeurIPS* (F. Pereira, C. Burges, L. Bottou, and K. Weinberger, eds.), vol. 25, Curran Associates, Inc., 2012.

[89] M. Arastuie, S. Paul, and K. S. Xu, "Chip: A hawkes process model for continuous-time networks with scalable and consistent estimation," 2019.

[90] S. Delattre, N. Fournier, and M. Hoffmann, "Hawkes processes on large networks," *The Annals of Applied Probability*, vol. 26, no. 1, pp. 216 – 261, 2016.

[91] X. Fan, B. Li, F. Zhou, and S. SIsson, "Continuous-time edge modelling using non-parametric point processes," *NeurIPS*, vol. 34, pp. 2319–2330, 2021.

[92] K. Ishiguro, T. Iwata, N. Ueda, and J. Tenenbaum, "Dynamic infinite relational model for time-varying relational data analysis," *NeurIPS*, vol. 23, 2010.

[93] T. Herlau, M. Mørup, and M. Schmidt, "Modeling temporal evolution and multiscale structure in networks," in *ICML*, pp. 960–968, 2013.

[94] C. Heaukulani and Z. Ghahramani, "Dynamic probabilistic models for latent feature propagation in social networks," in *ICML*, pp. 275–283, 2013.

[95] D. Durante and D. Dunson, "Bayesian Logistic Gaussian Process Models for Dynamic Networks," in *AISTATS*, vol. 33, pp. 194–201, 2014.

[96] D. Durante and D. B. Dunson, "Locally adaptive dynamic networks," *The Annals of Applied Statistics*, vol. 10, no. 4, pp. 2203–2232, 2016.

[97] R. Trivedi, M. Farajtabar, P. Biswal, and H. Zha, "Dyrep: Learning representations over dynamic graphs," in *ICLR*, 2019.

[98] E. Rossi, B. Chamberlain, F. Frasca, D. Eynard, F. Monti, and M. M. Bronstein, "Temporal graph networks for deep learning on dynamic graphs," *ICML 2020 Workshop*, 2020.

[99] N. Nakis, A. Çelikkanat, S. Lehmann, and M. Mørup, "A hierarchical block distance model for ultra low-dimensional graph representations," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–14, 2023.

[100] N. Nakis, A. Çelikkanat, and M. Mørup, "A hybrid membership latent distance model for unsigned and signed integer weighted networks," 2023.

[101] I. Chami, S. Abu-El-Haija, B. Perozzi, C. Ré, and K. Murphy, "Machine learning on graphs: A model and comprehensive taxonomy," 2022.

[102] M. Zitnik, M. Agrawal, and J. Leskovec, "Modeling polypharmacy side effects with graph convolutional networks," *Bioinformatics*, vol. 34, pp. i457–i466, jun 2018.

[103] R. Ying, R. He, K. Chen, P. Eksombatchai, W. L. Hamilton, and J. Leskovec, "Graph convolutional neural networks for web-scale recommender systems," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, KDD '18, (New York, NY, USA), p. 974–983, Association for Computing Machinery, 2018.

[104] P. Sen, G. Namata, M. Bilgic, L. Getoor, B. Gallagher, and T. Eliassi-Rad, "Collective classification in network data," in *The AI Magazine*, 2008.

[105] J. Xu and Y. Li, "Discovering disease-genes by topological features in human protein–protein interaction network," *Bioinformatics*, vol. 22, pp. 2800–2805, 09 2006.

[106] S. Xiao, S. Wang, Y. Dai, and W. Guo, "Graph neural networks in node classification: Survey and evaluation," *Mach. Vision Appl.*, vol. 33, jan 2022.

[107] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *J. Stat. Mech. Theory Exp.*, vol. 2008, no. 10, p. P10008, 2008.

[108] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical Review E*, vol. 69, feb 2004.

[109] P. Bedi and C. Sharma, "Community detection in social networks," *WIREs Data Mining and Knowledge Discovery*, vol. 6, no. 3, pp. 115–135, 2016.

[110] I.-C. Morarescu and A. Girard, "Opinion dynamics with decaying confidence: Application to community detection in graphs," *IEEE Transactions on Automatic Control*, vol. 56, no. 8, pp. 1862–1873, 2011.

[111] H. Mahmoud, F. Masulli, S. Rovetta, and G. Russo, "Community detection in protein-protein interaction networks using spectral and graph approaches," in *Computational Intelligence Methods for Bioinformatics and Biostatistics* (E. Formenti, R. Tagliaferri, and E. Wit, eds.), (Cham), pp. 62–75, Springer International Publishing, 2014.

[112] M. Salathé and J. H. Jones, "Dynamics and control of diseases in networks with community structure," *PLOS Computational Biology*, vol. 6, pp. 1–11, 04 2010.

[113] A. Karataş and S. Şahin, "Application areas of community detection: A review," in *2018 International Congress on Big Data, Deep Learning and Fighting Cyber Terrorism (IBIGDELFT)*, pp. 65–70, 2018.

[114] P. D. Hoff, "Modeling homophily and stochastic equivalence in symmetric relational data," in *NIPS*, p. 657–664, 2007.

[115] M. Nickel and D. Kiela, "Poincaré embeddings for learning hierarchical representations," 2017.

[116] M. Handcock, A. Raftery, and J. Tantrum, "Model-based clustering for social networks," *J. R. Stat. Soc.*, vol. 170, pp. 301 – 354, 2007.

[117] P. D. Hoff, "Bilinear mixed-effects models for dyadic data," *JASA*, vol. 100, no. 469, pp. 286–295, 2005.

[118] B. Karrer and M. E. Newman, "Stochastic blockmodels and community structure in networks," *Physical review E*, vol. 83, no. 1, p. 016107, 2011.

[119] T. Herlau, M. N. Schmidt, and M. Mørup, "Infinite-degree-corrected stochastic block model," *Physical review E*, vol. 90, no. 3, p. 032819, 2014.

[120] D. K. Wind and M. Mørup, "Link prediction in weighted networks," in *2012 IEEE International Workshop on Machine Learning for Signal Processing*, pp. 1–6, 2012.

[121] S. V. Beentjes and A. Khamseh, "Higher-order interactions in statistical physics and machine learning: A model-independent solution to the inverse problem at equilibrium," *Phys. Rev. E*, vol. 102, p. 053314, Nov 2020.

[122] R. Muolo, L. Gallo, V. Latora, M. Frasca, and T. Carletti, "Turing patterns in systems with high-order interactions," *Chaos, Solitons & Fractals*, vol. 166, p. 112912, 2023.

[123] J. Leskovec and A. Krevl, "SNAP Datasets: Stanford large network dataset collection," June 2014.

[124] H. C. White, S. A. Boorman, and R. L. Breiger, "Social structure from multiple networks. i. blockmodels of roles and positions," *American journal of sociology*, vol. 81, no. 4, 1976.

[125] P. W. Holland, K. B. Laskey, and S. Leinhardt, "Stochastic blockmodels: First steps," *Social networks*, vol. 5, no. 2, pp. 109–137, 1983.

[126] K. Nowicki and T. A. B. Snijders, "Estimation and prediction for stochastic blockstructures," *JASA*, vol. 96, no. 455, pp. 1077–1087, 2001.

[127] S. S. Epp, *Discrete Mathematics with Applications*. USA: Brooks/Cole, 4th ed., 2010.

[128] H. Tsutsu and Y. Morikawa, "An lp norm minimization using auxiliary function for compressed sensing," in *Proc. Int. Multiconf. Comp. Sci. Inf. Technol.*, 2012.

[129] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A K-Means clustering algorithm," *Applied Statistics*, vol. 28, no. 1, pp. 100–108, 1979.

[130] A. Tsitsulin, D. Mottin, P. Karras, and E. Müller, "VERSE," in *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*, ACM Press, 2018.

[131] X. Mao, P. Sarkar, and D. Chakrabarti, "On mixed memberships and symmetric nonnegative matrix factorizations," in *ICML*, vol. 70, 2017.

[132] K. Huang, N. D. Sidiropoulos, and A. Swami, "Non-negative matrix factorization revisited: Uniqueness and algorithm for symmetric decomposition," *IEEE Trans. Signal Process*, vol. 62, no. 1, pp. 211–224, 2014.

[133] A. Cutler and L. Breiman, "Archetypal analysis," *Technometrics*, vol. 36, no. 4, pp. 338–347, 1994.

[134] M. Mørup and L. Kai Hansen, "Archetypal analysis for machine learning," in *Workshop on Machine Learning for Signal Processing*, pp. 172–177, 2010.

[135] J. G. Skellam, "The frequency distribution of the difference between two poisson variates belonging to different populations.," *Journal of the Royal Statistical Society. Series A (General)*, vol. 109, no. Pt 3, pp. 296–296, 1946.

[136] L. Zhuang, C.-H. Lin, M. A. Figueiredo, and J. M. Bioucas-Dias, "Regularization parameter selection in minimum volume hyperspectral unmixing," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 12, pp. 9858–9877, 2019.

[137] B. Büeler, A. Enge, and K. Fukuda, "Exact volume computation for polytopes: a practical study," in *Polytopes—combinatorics and computation*, pp. 131–154, Springer, 2000.

[138] O. Shoval, H. Sheftel, G. Shinar, Y. Hart, O. Ramote, A. Mayo, E. Dekel, K. Kavanagh, and U. Alon, "Evolutionary trade-offs, pareto optimality, and the geometry of phenotype space," *Science*, vol. 336, no. 6085, pp. 1157–1160, 2012.

[139] Y. Hart, H. Sheftel, J. Hausser, P. Szekely, N. B. Ben-Moshe, Y. Korem, A. Tendler, A. E. Mayo, and U. Alon, "Inferring biological tasks using pareto analysis of high-dimensional data," *Nature methods*, vol. 12, no. 3, pp. 233–235, 2015.

[140] M. A. Hasan, J. Neville, and N. Ahmed, "Network sampling," in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, (New York, NY, USA), p. 1528, Association for Computing Machinery, 2013.

[141] A. E. Raftery, X. Niu, P. D. Hoff, and K. Y. Yeung, "Fast inference for the latent space network model using a case-control approximate likelihood," *J Comput Graph Stat.*, vol. 21, no. 4, pp. 901–919, 2012.

[142] J. A. González, F. J. Rodríguez-Cortés, O. Cronie, and J. Mateu, "Spatio-temporal point process statistics: A review," *Spatial Statistics*, vol. 18, pp. 505–544, 2016.

[143] R. L. Streit, *Poisson point processes: imaging, tracking, and sensing*. Springer Science & Business Media, 2010.

[144] D. Wang, C. Song, and A.-L. Barabási, "Quantifying long-term scientific impact," *Science*, vol. 342, no. 6154, pp. 127–132, 2013.

[145] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017.

[146] J. Yang and J. Leskovec, "Defining and evaluating network communities based on ground-truth," *Knowledge and Information Systems*, vol. 42, pp. 181–213, Jan 2015.

[147] S. Chacon, "2009 github challenge," 2009.

[148] J. Leskovec, J. Kleinberg, and C. Faloutsos, "Graph evolution: Densification and shrinking diameters," *ACM Trans. Knowl. Discov. Data*, vol. 1, no. 1, 2007.

[149] J. Leskovec and J. J. Mcauley, "Learning to discover social circles in ego networks," in *NIPS*, pp. 539–547, 2012.

[150] J. Leskovec, D. Huttenlocher, and J. Kleinberg, "Predicting positive and negative links in online social networks," in *WWW*, p. 641–650, 2010.

[151] R. West, H. S. Paskov, J. Leskovec, and C. Potts, "Exploiting social network structure for person-to-person sentiment analysis," *TACL*, vol. 2, pp. 297–310, 2014.

[152] S. Kumar, W. L. Hamilton, J. Leskovec, and D. Jurafsky, "Community interaction and conflict on the web," in *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, pp. 933–943, International World Wide Web Conferences Steering Committee, 2018.

[153] B. Ordozgoiti, A. Matakos, and A. Gionis, "Finding large balanced subgraphs in signed networks," in *Proceedings of The Web Conference 2020*, p. 1378–1388, 2020.

[154] T. Derr, C. Johnson, Y. Chang, and J. Tang, "Balance in signed bipartite networks," in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, (New York, NY, USA), p. 1221–1230, Association for Computing Machinery, 2019.

[155] M. E. Newman, "The structure of scientific collaboration networks," *Proceedings of the national academy of sciences*, vol. 98, no. 2, pp. 404–409, 2001.

[156] M. E. Newman, "Scientific collaboration networks. i. network construction and fundamental results," *Physical review E*, vol. 64, no. 1, p. 016131, 2001.

[157] M. E. Newman, "Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality," *Physical review E*, vol. 64, no. 1, p. 016132, 2001.

[158] F. Davletov, A. S. Aydin, and A. Cakmak, "High impact academic paper prediction using temporal and topological features," in *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, (New York, NY, USA), p. 491–498, Association for Computing Machinery, 2014.

[159] M. Singh, V. Patidar, S. Kumar, T. Chakraborty, A. Mukherjee, and P. Goyal, "The role of citation context in predicting long-term citation profiles: An experimental study based on a massive bibliographic text dataset," in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, (New York, NY, USA), p. 1271–1280, Association for Computing Machinery, 2015.

[160] H. S. Bhat, L.-H. Huang, S. Rodriguez, R. Dale, and E. Heit, "Citation prediction using diverse features," in *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, pp. 589–596, 2015.

[161] N. Shibata, Y. Kajikawa, and I. Sakata, "Link prediction in citation networks," *Journal of the American Society for Information Science and Technology*, vol. 63, no. 1, pp. 78–85, 2012.

[162] R. Yan, J. Tang, X. Liu, D. Shan, and X. Li, "Citation count prediction: Learning to estimate future citations for literature," pp. 1247–1252, 10 2011.

[163] A. Ibáñez, P. Larranaga, and C. Bielza, "Predicting citation count of bioinformatics papers within four years of publication," *Bioinformatics (Oxford, England)*, vol. 25, pp. 3303–9, 10 2009.

[164] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015.

[165] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[166] J. Anderson and D. Gerbing, "Structural equation modeling in practice: A review and recommended two-step approach," *Psychological bulletin*, vol. 103, pp. 411–423, May 1988.

[167] C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, may 2011.

[168] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," 2014.

[169] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing, "Mixed membership stochastic blockmodels," *J Mach Learn Res*, vol. 9, no. 65, pp. 1981–2014, 2008.

[170] S. Chanpuriya, C. Musco, K. Sotiropoulos, and C. E. Tsourakakis, "Node embeddings and exact low-rank representations of complex networks," *CoRR*, vol. abs/2006.05592, 2020.

[171] M. D. Ben Chamberlain and J. Clough, "Neural embeddings of graphs in hyperbolic space," in *MLG Workshop*, 2017.

[172] M. Nickel and D. Kiela, "Poincaré embeddings for learning hierarchical representations," in *NIPS*, vol. 30, 2017.

[173] A. Clauset, C. Moore, and M. E. Newman, "Hierarchical structure and the prediction of missing links in networks," *Nature*, vol. 453, no. 7191, pp. 98–101, 2008.

[174] D. M. Roy, C. Kemp, V. Mansinghka, and J. Tenenbaum, "Learning annotated hierarchies from relational data," in *NIPS*, vol. 19, 2007.

[175] T. Herlau, M. Mørup, M. N. Schmidt, and L. K. Hansen, "Detecting hierarchical structure in networks," in *CIP*, pp. 1–6, IEEE, 2012.

[176] C. Blundell and Y. W. Teh, "Bayesian hierarchical community discovery," in *NIPS*, vol. 26, 2013.

[177] T. P. Peixoto, "Hierarchical block structures and high-resolution model selection in large networks," *Physical Review X*, vol. 4, no. 1, 2014.

Appendix

Here, we are providing the complete papers that constitute this thesis. Due to copyright reasons, for papers not published under open access, we have included the accepted version instead of the published one.

A Hierarchical Block Distance Model for Ultra Low-Dimensional Graph Representations

Nikolaos Nakis, Abdulkadir Çelikkanat, Sune Lehmann, Morten Mørup

Abstract—Graph Representation Learning (GRL) has become central for characterizing structures of complex networks and performing tasks such as link prediction, node classification, network reconstruction, and community detection. Whereas numerous generative GRL models have been proposed, many approaches have prohibitive computational requirements hampering large-scale network analysis, fewer are able to explicitly account for structure emerging at multiple scales, and only a few explicitly respect important network properties such as homophily and transitivity. This paper proposes a novel scalable graph representation learning method named the Hierarchical Block Distance Model (HBDM). The HBDM imposes a multiscale block structure akin to stochastic block modeling (SBM) and accounts for homophily and transitivity by accurately approximating the latent distance model (LDM) throughout the inferred hierarchy. The HBDM naturally accommodates unipartite, directed, and bipartite networks whereas the hierarchy is designed to ensure linearithmic time and space complexity enabling the analysis of very large-scale networks. We evaluate the performance of the HBDM on massive networks consisting of millions of nodes. Importantly, we find that the proposed HBDM framework significantly outperforms recent scalable approaches in all considered downstream tasks. Surprisingly, we observe superior performance even imposing ultra-low two-dimensional embeddings facilitating accurate direct and hierarchical-aware network visualization and interpretation.

Index Terms—Latent Space Modeling, Complex Networks, Graph Representation Learning.

1 Introduction

Networks naturally arise in a plethora of scientific areas to model interactions between entities from physics to sociology and biology, with many instances such as collaboration, protein-protein interaction, and brain connectivity networks [1] to mention but a few. In recent years, Graph Representation Learning (GRL) approaches have attracted great interest due to their outstanding performance compared to classical techniques for arduous problems such as link prediction [2], node classification [3], [4], and community detection [5].

Numerous GRL methods have been proposed, see also [6] for a survey. The leading initial works are the random walk-based methods [4], [7]–[10], leveraging the Skip-Gram algorithm [11] to learn the node representations. Matrix factorization-based algorithms [6], [12] have also become prominent, extracting the embedding vectors by decomposing a designed feature matrix. Furthermore, neural network models [6], [13] have been proposed for graph-structured data, returning outstanding performance by combining node attributes and network structure when learning embeddings. Recent studies [14] aim to alleviate the computational burden of these algorithms through matrix sparsification tools [15], hierarchical representations [16], [17], or by fast hashing schemes [18].

Latent Space Models (LSMs) for the representation of graphs have been quite popular over the past years [19]–[25], especially for social networks analysis [26], [27] facilitating community extraction [28] and characterization of network polarization [29]. LSMs utilize the generalized linear model framework to obtain informative latent node embeddings while preserving network characteristics. The choice of latent effects in modeling the link probabilities between the nodes leads to different expressive capabilities characterizing network structure. In particular, in the Latent Distance Model (LDM) [30] nodes are placed closer in the latent space if they are similar or vice-versa. LDM obeys the triangle inequality and thus naturally represents transitivity [31], [32] (*"a friend of a friend is a friend"*) and network homophily [33], [34] (*a tendency where similar nodes are more likely to connect to each other than dissimilar ones*). Homophily is a very well-known and well-studied effect appearing in social networks [31], [33], [34] and essentially describes the tendency for people to form connections with those that share similarities with themselves. Similarities can be drawn from meta-data (observed node attributes) and may refer to shared demographic properties, political opinions, etc. Homophily has been observed among a broad range of collaborations (see [32] for a complete overview). Homophily can also be accounted for based on the unobserved attributes as defined by the LDM as shown in [35]. Homophily explains prominent patterns as expressed in social networks in terms of transitivity, as well as, balance theory ("the enemy of my friend is an enemy") [36]. More specifically, in an LDM we can extend the meaning of similarity to some unobserved (latent) covariates, i.e., latent embeddings $\mathbf{Z}$. The higher similarity between nodes translates here to a stronger relationship between the nodes and thereby a higher probability of observing connections. As a result, for two similar nodes $\{i, j\}$ the pairwise distance $|\mathbf{z}_i - \mathbf{z}_j|_2$ should be small which further implies that for a different node $\{k\}$ we obtain $|\mathbf{z}_i - \mathbf{z}_k|_2 \approx |\mathbf{z}_j - \mathbf{z}_k|_2$. The latter concludes that nodes $\{i, j\}$ are similar since they share similar relationships with the rest of the nodes.

The approach has been extended to bipartite networks in [37] by introducing mode-specific embedding vectors and community detection by endowing the LDM with a Gaus-

sian Mixture Model prior to promoting cluster structures in the latent space forming the latent position clustering model (LPCM) [35], [38]. However, the LDM is unable to account for possible stochastic equivalence as defined by the Stochastic Blockmodels [39], [40], i.e. (*"groups of nodes defined by shared intra- and inter-group relationships"*) defining non positive-semi-definite latent representations. The LSMs were advanced to characterize such stochastic equivalence by imposing an Eigenmodel admitting negative eigenvalues [41]. These latent space methods are attractive due to their simplicity, as they define well-structured inference problems and are characterized by high explanatory power [25]. The time and space complexities are their main drawbacks as the likelihood function scales by the number of node pairs (i.e., quadratically in the number of nodes for a unipartite graph) typically addressed using subsampling strategies [42].

Many real-world networks are composed of structures emerging at multiple scales which can be expressed using hierarchical representations [43]. Several methods have thus been advanced to such hierarchical representations including stochastic block model approaches [44]–[49] as well as agglomerative [50]–[52] and recursive partitioning [53] procedures relying on various measures of similarity. Importantly, learning node representations characterizing structure at multiple scales of the network can facilitate network visualization and the understanding of the inner dynamics of networks. Hierarchical representation of bipartite networks is of special interest due to the fact that most unipartite hierarchical clustering algorithms do not generalize to the bipartite case beyond clustering each mode separately or transforming the bipartite network into a unipartite representation. In the work of [54], the authors used the spectral partitioning algorithm of [55] and then applied k-means on the spectral space to get initial bi-clusters which were followed by divisive bi-splits to create a dendrogram. In this case, the spectral embedding space was not constructed to reflect explicitly the clustering criterion. In addition to divisive procedures, agglomerative clustering has also been proposed for bipartite networks. In the work of [56] a multi-objective function was designed and combined with classical community construction algorithms. One limitation here is that the network should be transformed into a unipartite structure.

Despite the many advantages of hierarchical structures and block models, one major limitation remains to accurately account for homophily [41], which is a key characteristic of social networks. More specifically, block models have been extended to explicitly impose a community structure [57], [58] but notable this only provides within-cluster homogeneity and thus homophily-like properties for the community relative to the other communities but not a hierarchy complying with such a structure. Whereas LPCM accounts for homophily it does not account for hierarchical structures and cluster structures are not strictly imposed beyond a prior to promoting the latent positions to form groups.

In this work, we propose a novel node representation learning approach, the Hierarchical Block Distance Model (HBDM)[1], as a reconciliation between hierarchical block structures of different scales and network properties such as homophily and transitivity. In particular, we propose a framework combining embedding and hierarchical characterization for graph representation learning. Importantly, we design a hierarchical structure that respects a linearithmic total time and space complexity, in terms of the number of nodes (i.e., $\mathcal{O}(N \log N)$), and at the same time provides an accurate interpretable representation of structure at different scales. Using the HBDM, we embed moderate-sized and large-scale networks containing more than a million nodes and establish the performance of our model in terms of link prediction and node classification to existing prominent graph embedding approaches. We further highlight how the inferred hierarchical organization can facilitate accurate visualization of network structure even when using only $D = 2$ dimensional representations providing favorable performance in all the considered GRL tasks; link prediction, node classification, and network reconstruction. Additionally, we show how our proposed framework extends the hierarchical multi-resolution structure to bipartite networks and provides the characterization of communities at multiple scales.

[1] *For implementation details please visit: github.com/Nicknakis/HBDM.*

2 The Hierarchical Block Distance Model

We first concentrate our study on undirected networks and later generalize our approach to bipartite graphs. We now provide the necessary definitions required throughout the paper. Let $\mathcal{G} = (V, E)$ be a graph where $N := |V|$ is the number of nodes and $Y_{N\times N} = (y_{i,j}) \in \{0,1\}^{N\times N}$ be the adjacency matrix of the graph such that $y_{i,j} = 1$ if the pair $(i, j) \in E$ otherwise it is equal 0, for all $1 \leq i < j \leq N$. We denote the latent representations of nodes by $\mathbf{Z} = (z_{i,d}) \in \mathbb{R}^{N\times D}$ where each row vector, $\mathbf{z}_i \in \mathbb{R}^D$, indicates the corresponding embedding of node $i \in \mathcal{V}$ in a D-dimensional space.

The most natural choice for modeling homophily and transitivity can be found in the Latent Space Model (LSM) which defines an $\mathbb{R}^D$-dimensional latent space in which every node of the graph is characterized through the unobserved but informative node-specific variables $\{\mathbf{z}_i \in \mathbb{R}^D\}$. These variables are considered sufficient to describe and explain the underlying relationships between the nodes of the network. The probability of an edge occurring is considered conditionally independent given the unobserved latent positions. Consequently, the total probability distribution of the network can be written as:

$$P(Y|\boldsymbol{Z}, \boldsymbol{\theta}) = \prod_{i<j}^{N} p(y_{i,j}|\mathbf{z}_i, \mathbf{z}_j, \boldsymbol{\theta}_{i,j}), \tag{1}$$

where $\boldsymbol{\theta}$ denotes any potential additional parameters, such as covariate regressors. A popular and convenient parameterization of Equation (1) for binary data is through the logistic regression model [30], [35], [59], [60]. In contrast, we adopt the Poisson regression model similar to [60] under a generalized linear model framework for the LSM. The use of a Poisson likelihood for modeling binary relationships in a network does not decrease the predictive performance nor the ability of the model to detect the network structure, as shown in [61]. It also generalizes the analysis to integer-weighted graphs. In addition, the exchange of the *logit* to a

log link function when transitioning from a Bernoulli to a Poisson model defines nice decoupling properties over the predictor variables in the likelihood [62], [63]. Utilizing the Poisson Latent Distance Model (LDM) of the LSM family framework, the rate of an occurring edge depends on a distance metric between the latent positions of the two nodes. In our formulation, we consider the LDM Poisson rate with node-specific biases or random-effects [35], [60] such that the expression for the Poisson rate becomes:

$$\lambda_{ij} = \exp\left(\gamma_i + \gamma_j - d(\mathbf{z}_i, \mathbf{z}_j)\right), \tag{2}$$

where $\gamma_i \in \mathbb{R}$ denotes the node-specific random-effects and $d_{ij}(\cdot,\cdot)$ denotes any distance metric obeying the triangle inequality $\{d_{ij} \leq d_{ik} + d_{kj},\ \forall (i,j,k) \in V^3\}$. Considering variables $\{\mathbf{z}_i\}_{i\in V}$ as the latent characteristics, Equation (2) shows that similar nodes will be placed closer in the latent space, yielding a high probability of an occurring edge and thus modeling homophily and satisfies network transitivity and reciprocity through the triangle inequality whereas the node-specific bias can account for degree heterogeneity. The conventional LDM rate utilizing a global bias, γ^g, corresponds to the special case in which $\gamma_i = \gamma_j = 0.5\gamma^g$. As in [30], we presently adopt the Euclidean distance as the choice for the distance metric $d_{ij}(\cdot,\cdot)$.

2.1 Designing A Linearithmic Complexity

Our goal is to design a Hierarchical Block Model preserving homophily and transitivity properties with a total complexity allowing for the analysis of large-scale networks. Our HBDM, defines the rate of a link between each network dyad $\{i,j\} \in V \times V$ based on the Euclidean distance, as shown in Equation (2). Therefore, we can define a block-alike hierarchical structure by a divisive clustering procedure over the latent variables in the Euclidean space. The total optimization cost of such a model though should have a linearithmic upper bound complexity to make large-scale analysis feasible. Introducing a number of clusters K equal to the number of nodes N in the HBDM, leads to the same log-likelihood as of the standard LDM, defining a sum over each ordered pair of the network, as:

$$\begin{aligned}\log P(Y|\mathbf{\Lambda}) &= \sum_{i<j}\Big(y_{ij}\log(\lambda_{ij}) - \lambda_{ij}\Big) \\ &= \sum_{i<j:y_{ij}=1}\log(\lambda_{ij}) \;-\; \sum_{i<j}\lambda_{ij}\,,\end{aligned} \tag{3}$$

For brevity, we presently ignore the linear scaling by dimensionality D of the above log-likelihood function. Notably, the link contribution $\sum_{y_{i,j}=1}\log(\lambda_{i,j})$ is responsible for positioning "similar" nodes closer in the latent space, expressing the desired homophily.

In addition, large networks are highly sparse [64] with the number of edges being proportional to the number of nodes in the network. As a result, the computation of the link contribution is relatively cheap, scaling linearithmic or sub-linearithmic (as shown in supplementary). Most importantly, the link term removes rotational ambiguity between the different blocks of the hierarchy (as discussed later). For these three reasons, no block structure is imposed on the calculation of the link contribution. The second term acts as the repelling force for dissimilar nodes and requires the computation of all node pairs scaling as $\mathcal{O}(N^2)$ making the evaluation of the above likelihood infeasible for large networks. By enforcing a block structure, i.e., akin to stochastic block models [39], [40], [65], when grouping the nodes into K clusters we define the rate between block k and k' in terms of their distance between centroids. A simple block structure without a hierarchy would lead to the following non-link expression:

$$\begin{aligned}\sum_{i<j}\lambda_{ij} \approx \sum_{k=1}^{K}\Bigg(&\sum_{\substack{i<j\\ i,j\in C_k}}\exp\big(\gamma_i+\gamma_j-||\mathbf{z}_i-\mathbf{z}_j||_2\big) \\ &+\sum_{k'>k}^{K}\sum_{i\in C_k}\sum_{j\in C_{k'}}\exp\left(\gamma_i+\gamma_j-||\boldsymbol{\mu}_k-\boldsymbol{\mu}_{k'}||_2\right)\Bigg),\end{aligned} \tag{4}$$

where $\boldsymbol{\mu}_k$ denotes the k'th cluster centroid of the set $\boldsymbol{C} = \{C_1,\ldots,C_K\}$, and has absorbed the dependency over the variables $\boldsymbol{Z} \in \mathbb{R}^{N\times D}$. More specifically, the cluster centroids $\boldsymbol{\mu}_k$ are implicit parameters defined as a function over the latent variables, as we will show later. Overall, considering the principle that connected and homophilic nodes will be placed closer in the latent space, this expression generalizes this principle by introducing a clustering procedure that obeys "cluster-homophily" and "cluster-transitivity" over the latent space. More specifically, we can assume that closely related nodes will be positioned in the same cluster while related or interconnected clusters will also be positioned close in the latent space, providing an accurate block structure schema. As opposed to the LPCM where clustering structures are imposed through a prior, the above formulation strictly defines the clustering structure as shared overall proximity between blocks as defined by the distances between centroids of the formed groups.

2.1.1 A Hierarchical Representation

In order to obtain the desired hierarchical representation, we define hierarchical clustering via a divisive procedure. In detail, we organize the embedded clusters into a hierarchy using a tree structure, defining a cluster dendrogram. The root of the tree is a single cluster containing the total amount of latent variable embeddings $\boldsymbol{Z}$. At every level of the tree, we perform partitioning until we obtain leaf nodes containing equal or less than the desired number of nodes, N_{leaf}. This number is chosen with respect to our linearithmic complexity upper bound and set as $N_{\text{leaf}} = \log N$, resulting in approximately $K = N/log(N)$ total clusters. The tree-nodes belonging to a specific tree-level are considered the clusters for that specific tree height. Every novel partition of a non-leaf node is performed only on the set of points allocated to the parent tree-node (cluster). For every level of the tree, we consider the pairwise distances of datapoints belonging to different tree-nodes as the distance between the corresponding cluster centroids, as illustrated by Fig. 1 (ii). Based on these distances, we calculate the likelihood contribution of the blocks and continue with binary splits, down the tree, for the non-leaf tree-nodes. When all tree-nodes are considered as leaves, we calculate analytically the inner cluster pairwise distances for the corresponding likelihood contribution of analytical blocks, as shown in the

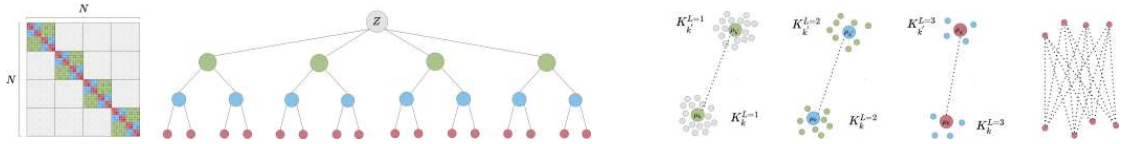


Fig. 1. Schematic representation of the distance matrix calculation for a hierarchical structure of the tree of height $L = 3$ and for the number of observations $N = 64$.

last part of Fig. 1 (ii). The latter analytical calculation comes at a linearithmic cost of $\mathcal{O}(KN_{\text{leaf}}^2) = \mathcal{O}(N \log N)$ while enforces the homophily and transitivity properties of the model since for the most similar nodes the HBDM behaves explicitly as the standard LDM.

We can thereby define a Hierarchical Block Distance Model with Random Effects (HBDM-RE) as:

$$\begin{aligned}\log P(Y|\mathbf{Z},\boldsymbol{\gamma}) = &\sum_{\substack{i<j \\ y_{i,j}=1}} \Big(\gamma_i + \gamma_j - ||\mathbf{z}_i - \mathbf{z}_j||_2\Big) \\ &- \sum_{k=1}^{K_L} \Bigg(\sum_{\substack{i<j \\ i,j \in C_k^{(L)}}} \exp\left(\gamma_i + \gamma_j - ||\mathbf{z}_i - \mathbf{z}_j||_2\right)\Bigg) \\ &- \sum_{l=1}^{L} \sum_{k=1}^{K_l} \sum_{k'>k}^{K_l} \Bigg(\exp\left(-||\boldsymbol{\mu}_k^{(l)} - \boldsymbol{\mu}_{k'}^{(l)}||_2\right) \\ &\times \Big(\sum_{i \in C_k^{(l)}} \exp\left(\gamma_i\right)\Big)\Big(\sum_{j \in C_{k'}^{(l)}} \exp\left(\gamma_j\right)\Big)\Bigg), \end{aligned} \tag{5}$$

where $l \in \{1, \ldots, L\}$ denotes the l'th dendrogram level, k_l is the index representing the cluster id for the different tree levels, and $\boldsymbol{\mu}_k^{(l)}$ the corresponding centroid. We also consider a Hierarchical Block Distance Model (HBDM) without the random effects setting $\gamma_i = 0.5\gamma^g$. For a multifurcating tree splitting in K clusters and having $N/log(N)$ terminal nodes (clusters), the number of internal nodes are $\mathcal{O}(N/(K \log N))$ and each node needs to evaluate $\mathcal{O}(K^2)$ pairs providing an overall complexity of $\mathcal{O}(NK/\log N)$, thus $K \leq \log N^2$ to achieve $\mathcal{O}(N \log N)$ scaling [66].

2.1.2 Divisive partitioning using k-means with a Euclidean distance metric

Whereas the likelihood in Equation (5) can be directly minimized by assigning nodes to the clusters given by the tree structure, this evaluation for all N nodes scales prohibitively as $\mathcal{O}(N^2/\log N)$. To reduce this scaling, we use a more efficient divisive partitioning procedure, minimizing the Euclidean norm $||\boldsymbol{\mu}_{kl} - \boldsymbol{\mu}_{k'_l}||_2$. The divisive clustering procedure thus relies on the following Euclidean norm objective

$$J(\mathbf{r}, \boldsymbol{\mu}) = \sum_{i=1}^{N} \sum_{k=1}^{K} r_{ik} ||\mathbf{z}_i - \boldsymbol{\mu}_k||_2, \tag{6}$$

where k denotes the cluster id, $\mathbf{z}_i$ is the i'th data observation, r_{ik} the cluster responsibility/assignment, and $\boldsymbol{\mu}_k$ the cluster centroid.

This objective function is unfortunately not accounted for by existing k-means clustering algorithms relying on the squared Euclidean norm. We therefore presently derive an optimization procedure for k-means clustering with Euclidean norm utilizing the auxiliary function framework of [67] developed in the context of compressed sensing. We define an auxiliary function for (6) as:

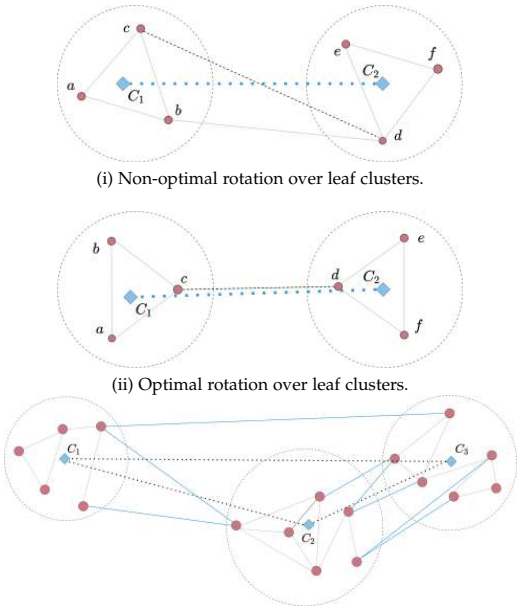
$$J^+(\boldsymbol{\phi}, \mathbf{r}, \boldsymbol{\mu}) = \sum_{i=1}^{N} \sum_{k=1}^{K} r_{ik} \left(\frac{||\mathbf{z}_i - \boldsymbol{\mu}_k||_2^2}{2\phi_{ik}} + \frac{1}{2}\phi_{ik} \right), \tag{7}$$

where $\boldsymbol{\phi}$ are the auxiliary variables. Thereby, minimizing Equation (7) with respect to ϕ_{nk} yields $\boldsymbol{\phi}_{ik}^* = ||\mathbf{z}_i - \boldsymbol{\mu}_k||_2$ and by plugging $\boldsymbol{\phi}_{ik}^*$ back to (6) we obtain $J^+(\boldsymbol{\phi}^*, \mathbf{r}, \boldsymbol{\mu}) = J(\mathbf{r}, \boldsymbol{\mu})$ verifying that (7) is indeed a valid auxiliary function for (6). The algorithm proceeds by optimizing cluster centroids as $\boldsymbol{\mu}_k = \left(\sum_{i \in k} \frac{\mathbf{z_i}}{\phi_{ik}} / \sum_{i \in k} \frac{1}{\phi_{ik}} \right)$ and assigning points to centroids as $\arg\min_{\boldsymbol{C}} = \sum_{k=1}^{K} \sum_{\boldsymbol{z} \in C_k} \left(\frac{||\boldsymbol{z} - \boldsymbol{\mu}_k||_2^2}{2\boldsymbol{\phi}_k} + \frac{1}{2}\boldsymbol{\phi}_k \right)$ upon which $\boldsymbol{\phi}_k$ is updated. The overall complexity of this procedure is $\mathcal{O}(TKND)$ [68] where T is the number of iterations required to converge. As shown in [67], Equation (7) is a special case of a general algorithm for an $l_p (0 < p < 2)$ norm minimization using an auxiliary function with the algorithm converging faster the smaller p is. For a detailed study of the efficiency of the optimization procedure under such an auxiliary function, see [67].

A simple approach to construct the tree structure would be to use the above Euclidean k-means procedure to split the nodes into $K = N/\log(N)$ clusters and construct the tree according to agglomeration as in hierarchical clustering. Unfortunately, such a strategy is computationally prohibitive. For that, in the coarser level (first layer of the tree), we choose to split to the maximally allowed clusters of $K = \log N$ allowing scaling of $\mathcal{O}(N \log N)$. It would be tempting to continue splitting into $\log N$ clusters, however, for a balanced multifurcating tree with $N/\log N$ leaf clusters, it will result in a height scaling as $\mathcal{O}(\log N/\log\log N)$ and thus an overall complexity of $\mathcal{O}(N \log^2(N)/\log\log N)$ [66]. Whereas a balanced binary tree at all levels below the root results in a height scaling as $\mathcal{O}(\log N)$ providing an overall complexity when including the linear scaling by dimensionality D of $\mathcal{O}(DN \log N)$ (as each level of the tree defines $\mathcal{O}(DN)$ operations). Fig. 1 (i), illustrates the resulting tree[2] for a small problem of $N = 64$ nodes in which we first split into 4 ($\approx \log(64)$) clusters and subsequently create binary splits until each leaf cluster contains 4 ($\approx \log(64)$) nodes.

2.1.3 Expressing Homophily and Transitivity

A central component to preserve homophily and transitivity of the HBDM is not to approximate the link terms at the level of the block as in (hierarchical) SBMs but to calculate analytically the link contribution of the log-likelihood across the total hierarchy beyond the leaf/analytical blocks. In Fig., 2 (i) and (ii), two leaf clusters are illustrated and connected with a link. Assume that we only calculate the distance inside the blocks analytically and that both the link and non-link contributions of pairs belonging to different clusters are approximated based on their centroids' distance. This essentially would allow for any rotation of each cluster for all clusters in the hierarchy since the inner-block distances (analytical), as well as the centroid distances, would not change by such rotations, yielding exactly the same likelihood (block-level rotational invariance). In that case, homophily would be violated as, e.g., the distance between nodes c and d would not necessarily be smaller than the distance with other inter-cluster pairs (ex: Fig. 2 (i)), showing that the rotation of the blocks substantially impacts the homophily properties of the HBDM. Calculating the link contributions between the different clusters analytically solves this ambiguity since the likelihood is penalized higher when nodes c and d are positioned in a non-rotational-aware way. The computational cost imposed by accounting for all the link terms analytically is that the model complexity depends on the number of edges of the network (a total block structure would strictly be linearithmic in complexity). Nevertheless, we show empirically in the supplementary that the number of network edges E scales linearly with $N \log N$ and thus this analytical term respects our complexity bound. In Fig. 2 (iii), we present clusters defining cases of block inter-connections of sparsely connected blocks ($\{C_1, C_3\}$, $\{C_1, C_2\}$) and densely connected blocks $\{C_2, C_3\}$. Whereas the analytical inter-cluster links (blue lines) are responsible for fixing the block rotation the inter-cluster links also drive the cluster-level proximities of centroids ensuring cluster homophily and transitivity.



(i) Non-optimal rotation over leaf clusters.

(ii) Optimal rotation over leaf clusters.

(iii) Three latent block structures.

Fig. 2. The clusters within the dashed circles denote the leaf block structures. The red circles and blue rhombuses indicate the node embeddings and the centroids, respectively. Gray lines represent the links and the dashed lines the distance between the cluster centers.

[2]For visualization purposes only, we show equally sized clusters.

Pairwise distances in the HBDM stays invariant to rotation, reflection, and translation of the latent space due to its LDM inheritance [30] these isometries can be resolved via a Singular-Value-Decomposition procedure as provided in the supplementary. Whereas the analytical link term calculations provide rotational awareness to the HBDM clusters, we continue by investigating the conditions in which a continuous operation defining infinitesimal rotations (with respect to the cluster centroid) is admissible leaving the loss function of Equation (5) invariant to continuous rotations. In Lemma 2.1 (proof given in the supplementary material), we start our investigation of this problem by showing that blocks with a unique inter-cluster link connection reduce the clusters' degree of rotational freedom by one.

Lemma 2.1. *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a graph and let $\mathcal{C}$ be a cluster with its centroid located at $\boldsymbol{\mu} \in \mathbb{R}^D$ having an edge $(i, j) \in \mathcal{E}$ for some $i \in \mathcal{C}$ and $j \in \mathcal{V} \backslash \mathcal{C}$ such that $\mathbf{z}_i \neq \boldsymbol{\mu}$. If $\tilde{\mathbf{z}}_i = \boldsymbol{\mu} + \mathbf{R}(\boldsymbol{\theta})(\mathbf{z}_i - \boldsymbol{\mu})$ such that $\mathbf{R}(\boldsymbol{\theta})$ is a rotation matrix acting on the embeddings of nodes in cluster $\mathcal{C}$, then the maximum degree of freedom of any infinitesimal λ_{ij}-invariant rotation is defined by $\boldsymbol{\theta} \in \mathbb{R}^{D-2}$.*

A direct consequence of Lemma 2.1 is that for a two-dimensional embedding, there is no possible continuous rotation of a cluster having only one external edge. Since there is a path from one node to all others in a connected graph, every cluster must have at least one external link. For the general case of blocks having multiple inter-cluster edges, rotations preserving the total sum of pairwise distances among node embeddings are highly unlikely, as discussed in the supplementary. Consequently, we can for connected networks expect uniqueness of a (local) minima solutions with no continuous admissible rotations leaving the HBDM loss function of Equation (5) invariant.

TABLE 1
Complexity analysis of methods. $N := |V|$ is the vertex set, $|E|$: edge set, $\mathcal{W}$: number of walks, $\mathcal{L}$: walk length, H: height of the hierarchical tree, D: node representation size, k: number of negative instances, q: order value, c: Chebyshev expansion order, γ: window size, α_1 and α_2 constants such as $\alpha_1, \alpha_2 \ll N$.

Method	Complexity
DEEPWALK	$\mathcal{O}\left(\gamma N \log\left(N\right) \mathcal{W} \mathcal{L} \mathcal{D}\right)$
NODE2VEC	$\mathcal{O}\left(\gamma N \mathcal{W} \mathcal{L} \mathcal{D} k\right)$
LINE	$\mathcal{O}\left(\lvert E \rvert D k\right)$
NETMF	$\mathcal{O}\left(N^2 D\right)$
NETSMF	$\mathcal{O}\left(\lvert E \rvert(\gamma + D) + N D^2 + D^3\right)$
RANDNE	$\mathcal{O}\left(N D^2 + \lvert E \rvert D q\right)$
LOUVAINNE	$\mathcal{O}\left(\lvert E \rvert \mathcal{H} + N D\right)$
PRONE	$\mathcal{O}\left(N D^2 + \lvert E \rvert c\right)$
VERSE	$\mathcal{O}\left(N(\mathcal{W} + k D)\right)$
HBDM	$\mathcal{O}\left(\alpha_2 N \log\left(N\right) D\right)$

2.1.4 Extension to Bipartite Networks

Our proposed frameworks, HBDM and HBDM-RE generalize to both directed and bipartite graphs. In the following, we provide the mathematical extension for the bipartite case (the directed network formulation of our proposed model can be considered a special case of the bipartite framework in which self-links are removed and thus omitted from the below log-likelihood). For a bipartite network with adjacency matrix $Y^{N_1 \times N_2}$ we can formulate the log-likelihood as:

$$\begin{aligned}\log P(Y|\mathbf{\Lambda}) = &\sum_{\substack{i<j \\ y_{i,j}=1}} \Big(\psi_i + \omega_j - ||\mathbf{w}_i - \mathbf{v}_j||_2\Big) \\ &- \sum_{k_L=1}^{K_L} \Bigg(\sum_{i,j \in C_{k_L}} \exp\left(\psi_i + \omega_j - ||\mathbf{w}_i - \mathbf{v}_j||_2\right)\Bigg) \\ &- \sum_{l=1}^{L} \sum_{k=1}^{K_l} \sum_{k'>k}^{K_l} \Bigg(\exp\left(-||\boldsymbol{\mu}_k^{(l)} - \boldsymbol{\mu}_{k'}^{(l)}||_2\right) \\ &\times \Big(\sum_{i \in C_k^{(l)}} \exp\left(\psi_i\right)\Big)\Big(\sum_{j \in C_{k'}^{(l)}} \exp\left(\omega_j\right)\Big)\Bigg), \end{aligned} \tag{8}$$

where $\{\boldsymbol{\mu}_k^{(l)}\}_{k=1}^{K_L}$ are the latent centroids which have absorbed the dependency of both sets of latent variables $\{\mathbf{w}_i, \mathbf{v}_j\}$ while we define the Poisson rate as:

$$\lambda_{ij} = \exp\big(\psi_i + \omega_j - d(\mathbf{w}_i, \mathbf{v}_j)\big), \tag{9}$$

where ψ_i and ω_j are the corresponding random effects and $\{\mathbf{w}_i,\ \mathbf{v}_j\}$ are the latent variables of the two disjoint sets of the vertex set of sizes N_1 and N_2, respectively. In this setting, we use our divisive Euclidean distance hierarchical clustering procedure over the concatenation $\mathbf{z} = [\mathbf{w}; \mathbf{v}]$ of the two sets of latent variables. Therefore, we define an accurate hierarchical block structure for bipartite networks, with each block including nodes from both of the two disjoint modes. Here, a centroid is considered a leaf if the corresponding tree-cluster contains less than $\log(N_1)$ of the latent variables $\{\mathbf{w}_i\}_{i=1}^{N_1}$ or less than $\log(N_2)$ of $\{\mathbf{v}\}_{j=1}^{N_2}$.

2.1.5 Complexity Comparison

TABLE 1 provides a comparison between time complexities of several prominent GRL methods in terms of their Big $\mathcal{O}$ notation, similar to [69]. We observe that our proposed HBDM is positioned as one of the most competitive frameworks. In terms of space complexity, our model defines a linearithmic complexity contrary to the majority of the considered baselines which are usually characterized by a quadratic space complexity [69]. (For a more detailed discussion please visit the supplementary.)

3 EXPERIMENTS

We extensively evaluate the performance of our method compared to baseline graph representation learning approaches on networks of various sizes and structures. We have conducted all the experiments regarding the HBDM and HBDM-RE on a 32 GB Tesla V100 GPU machine with 5120 CUDA cores, and a 1380 MHz clock. For the HBDM and HBDM-RE models, we optimize the negative log-likelihood via the Adam [70] optimizer with learning rate $lr \in [0.01, 0.1]$. For both frameworks, we build the hierarchical structure by running the k-means procedure every $t = 25$ iterations. Experiments regarding the baselines have been conducted on an Intel Xeon Gold 6342 CPU with 24 cores, 2800 MHz clock, and 512 GB memory. The implementation for HBDM and HBDM-RE is GPU-focused using PyTorch 1.12.1, exploiting parallel computations (running the frameworks on a CPU machine leads to substantially higher runtimes). We argue, that runtime comparison in terms of real-time is a biased estimate between different models since it correlates highly with the programming language, parallelization schemes, etc. For that, we instead compare theoretical complexities in terms of their Big $\mathcal{O}$ notation. In all TABLES, we denote with bold digits the best-performing score while we underline the second-best.

Datasets: We have performed the experiments on ten undirected networks of various sizes and structures: a citation network (*Cora* [71]), social interaction graphs (*Facebook* [72], *YouTube* [73], [74], *Flickr* [74], *Flixster* [75]), product-label network (*Amazon* [73]) and collaboration networks (*Dblp* [76], *AstroPh* [77], *GrQc* [77], *HepTh* [77]). Each network is considered as unweighted for the consistency of the experiments. The detailed statistics of the networks are provided by TABLE 2. All of the considered networks have been widely adopted and extensively used as benchmarks in the GRL literature [78].

Baseline Methods: In our experiments, we have run various graph representation learning methods in order to evaluate the performance of our approach. The prominent GRL frameworks used in this study are: (i) DEEPWALK [7], (ii) NODE2VEC [4], (iii) LINE [8], (iv) NETMF [79]. In addition, we consider five scalable graph embedding approaches: (v) NETSMF [15], (vi) RANDNE [18], (vii) PRONE [14], (viii) LOUVAINNE [16], (ix) VERSE [69]. For more details see the supplementary material. In our analysis, we considered GRAPHSAGE [13] as a prominent member of the family of Graph Neural Networks (GNNs). Our study focuses on the setting where node meta-data are not available. In such a setting, GRAPHSAGE was characterized by a close-to-random performance and thus not presented.

3.1 Link Prediction

We report results for the area under the curve of the receiver operator characteristic (AUC). For the experimental setup, we follow the commonly applied strategy [4], [7] and remove half of the edges while keeping the residual network connected. This strategy is not feasible for large-scale networks since checking if the residual network stays connected after each removed link results in a high runtime complexity. For that, we hide 30% of the edges for these networks and extract the giant component (after the link removal) which is treated as the residual network. Extensive details for the link prediction experiments, as well as, Precision-Recall AUC scores are given in the supplementary. Error bars across 5 re-runs for the following AUC scores were found to be on the scale of 10^{-3} and thus negligible.

Effectiveness and Efficiency of the Multi-Scale Approximation: In Fig. 3a, we provide an effectiveness analysis of the HBDM likelihood when contrasted with its full likelihood estimation evaluated on the moderate-sized network

TABLE 2
Statistics of undirected networks. N: number of nodes, $|E|$: number of edges.

	Cora	*Dblp*	*AstroPh*	*GrQc*	*Facebook*	*HepTh*	*Amazon*	*YouTube*	*Flickr*	*Flixster*
N	2,708	27,199	17,903	5,242	4,039	8,638	334,868	1,138,499	1,715,255	2,523,386
$\|E\|$	5,278	66,832	197,031	14,496	88,234	24,827	925,876	2,990,443	15,555,042	7,918,801

of Facebook (results for more networks are provided in the supplementary material). We here observe that the HBDM likelihood essentially approximates the true full likelihood providing systematically slightly lower likelihood estimates which we attribute to the small structural differences between calculating the distances analytically versus in a hierarchical block manner. A close approximation to the true likelihood provides evidence for multi-scale structures that characterize networks, yielding a high effectiveness of the HBDM framework. In addition, in Fig. 3a we see fluctuations in the likelihood which is an immediate result of building the network hierarchy from scratch every 25'th iteration. Importantly, despite the fact that k-means is notoriously known to be an NP-hard problem [80], [81], we observe that rebuilding the hierarchy has a minimum effect on the value of the likelihood, highlighting the stability of the inferred hierarchy in the HBDM. Furthermore, TABLE 3 conveys information about the comparison between an HBDM (approx) framework where all link distances are approximated by the centroid distances and the proposed HBDM where link distances are calculated analytically. We witness for the *Facebook* network how the rotational awareness induced by explicitly accounting for all links in the likelihood (as explained in subsection 2.1.3) increases the predictive capability of the model and thus its efficiency (similar results were obtained for all networks).

Moderate-Sized Networks: Results for the moderate-sized networks are given in TABLE 4. The symbol "-" indicates that the running time of the corresponding model takes more than 20 hours and "x" shows that the method is not able to run due to insufficient memory space. We observe that the HBDM and HBDM-RE perform significantly better or on par with the performance of the considered baseline approaches. In particular, the HBDM and HBDM-RE perform better than all the non-LDM baselines when $D = 2$. It highlights the superiority of LDMs in learning very low-dimensional network representations that accurately account for the network structure. We further observe that representing degree heterogeneity with random effects provides extended representational power as the HBDM-RE consistently outperforms the HBDM. Comparing our framework with the classic LDM-RE and LDM, we mostly see on-par results experimentally which we attribute to the hierarchical structure well-preserving properties of homophily and transitivity.

Large-Scale Networks: Results for the large-scale networks are given in TABLE 5. Again, we observe that HBDM-RE was on par with the most competitive baselines of NETSMF and VERSE while significantly outperforming the rest across networks. We also here find that the inclusion of random effects in the LDMs improves performance highlighting the importance of explicitly accounting for degree heterogeneity also for large networks. Notably, we again detect very good performance for the HBDM-RE, but also for NETSMF and VERSE when utilizing the very low embedding dimension of $D = 2$.

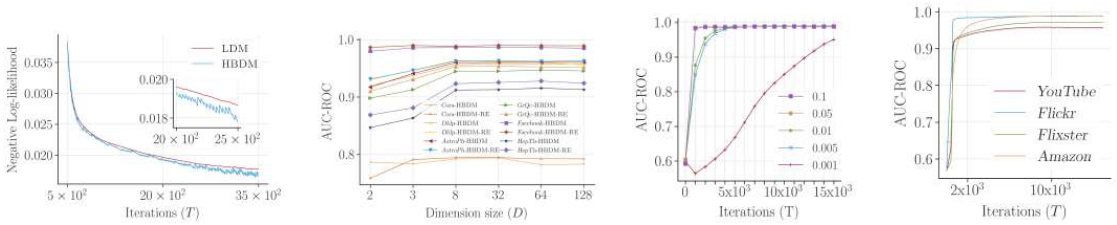
Bipartite Networks: We validate the performance of our proposed framework for bipartite structures by reporting the AUC score. We perform our experiments on three bipartite networks: (1) *Drug-Gene* [82] ($N_1 = 5,017$, $N_2 = 2,324$, $|E| = 15,138$), (2) *GitHub* [83] ($N_1 = 56,519$, $N_2 = 120,867$, $|E| = 440,237$), and (3) *Gottron-Reuters* [84] ($N_1 = 21,557$, $N_2 = 38,677$, $|E| = 1,464,182$) following the same experimental setting as in the undirected case of the moderate-sized networks (network details are given in the supplementary). We provide the results in TABLE 7 where we witness how the random-effects formulation of HBDM-RE and VERSE outperform all the baselines and in most cases significantly. Another interesting observation is that the random-effects model has a notably higher performance than the corresponding global bias model, as the three studied networks have a high degree of heterogeneity.

3.2 Hyperparemeter sensitivity

We here study the effect of hyperparameters introduced by the HBDM frameworks. Contrary to many GRL approaches, our models only define three hyperparameters which include the embedding dimensionality D, the number of training iterations, and the learning rate lr for the optimizer. In Fig. 3b, we view the predictive performance as a function of latent dimension D and here, in general, attain modest improvements in the predictive performance when increasing the embedding dimensions from $D = 2$ to $D = 8$ with no further improvements increasing to $D = 128$, highlighting the efficiency in which HBDM and HBDM-RE utilize very low-dimensional representations. Fig. 3c, shows the effect that the learning rate has on performance. We here witness that very small choices $lr \approx 0.001$ define a very slowly increasing performance. Medium magnitude choices of $lr \in [0.005, 0.01]$ define faster convergence while the optimum choices defining very rapid performance saturation exist in the $lr \in [0.05, 0.1]$ regime. In Fig. 3d, we investigate the convergence of the best performing HBDM-RE $D = 8$ for the large networks, and we witness that the

TABLE 3
AUC-ROC scores for varying dimension sizes on the *Facebook* network for a model approximating the link terms (top two rows) and for the proposed model which calculates analytically the link terms (bottom two rows).

Dimension (D)	2	3	8	32	64	128
HBDM (approx)	.656	.797	.946	.943	.940	.945
HBDM-RE (approx)	.802	.838	.909	.932	.940	.942
HBDM	.980	.986	.986	.987	.986	.985
HBDM-RE	.986	.990	.988	.989	.989	.989



(a) (b) (c) (d)

Fig. 3. (a) NLL comparison between HBDM and LDM for *Facebook* with $D = 2$. (b) AUC-ROC performance over various networks for varying embedding sizes. (c) Performance sensitivity over different learning rate choices for the optimizer in terms of AUC-ROC scores for the *Facebook* network. (d) AUC scores of HBDM-RE in terms of iterations sensitivity for large-scale networks.

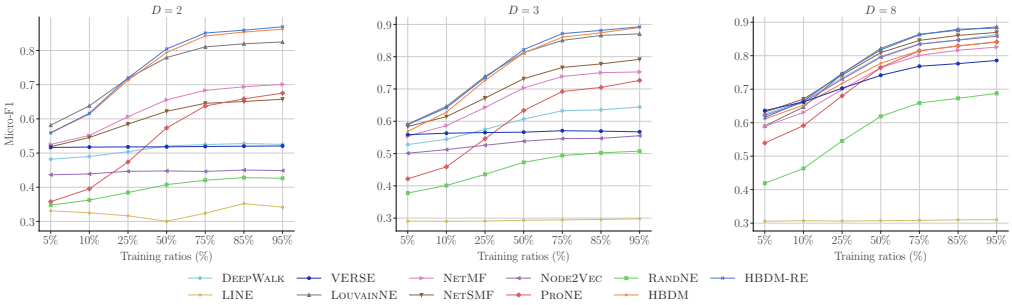


Fig. 4. Micro-F1 scores for the classification task considering different low-dimensions and training set ratios for the *DBLP* network.

TABLE 4
AUC scores for representation sizes of 2 and 8 over moderate-sized networks.

	AstroPh		*GrQc*		*Facebook*		*HepTh*		*Cora*		*DBLP*	
Dimension (D)	2	8	2	8	2	8	2	8	2	8	2	8
DeepWalk	.831	.945	.845	.919	.958	.986	.773	.874	.684	.782	.803	.939
Node2Vec	.825	.950	.809	.884	.914	**.988**	.780	.881	.640	.776	.803	.945
LINE	.632	.910	.688	.920	.751	.980	.659	.874	.634	.521	.625	.503
NetMF	.800	.814	.830	.860	.872	.935	.757	.792	.629	.739	.838	.858
NetSMF	.828	.891	.756	.805	.907	.976	.705	.810	.605	.737	.766	.857
RandNE	.524	.554	.534	.560	.614	.657	.519	.509	.508	.556	.508	.517
LouvainNE	.798	.813	.861	.868	.957	.958	.774	.874	.767	.747	.900	.904
ProNE	.768	.907	.818	.883	.900	.971	.678	.823	.675	.764	.813	.924
Verse	.899	**.974**	.885	.941	.970	**.992**	.844	.910	.749	.760	.910	.955
LDM	.925	x	.915	.943	.989	.991	.855	.919	.780	.786	.918	x
LDM-RE	**.943**	x	.925	.944	**.990**	**.992**	.869	.917	.770	.787	.926	x
HBDM	.920	.960	.917	.944	.980	.986	.853	.915	**.786**	.792	.919	.956
HBDM-RE	.939	.964	**.926**	**.953**	.986	.988	**.871**	**.924**	.774	**.795**	**.930**	**.963**

TABLE 5
AUC for varying representation sizes over the large-scale networks.

	YouTube			*Flickr*			*Flixster*			*Amazon*		
Dimension (D)	2	3	8	2	3	8	2	3	8	2	3	8
DeepWalk	.822	.891	.921	.889	.937	.972	.820	.866	.921	.839	.932	.972
Node2Vec	-	-	-	-	-	-	-	-	-	.813	.880	.968
LINE	.660	.832	.878	.685	.889	.812	.523	.868	.936	.626	.501	.500
NetMF	x	x	x	x	x	x	x	x	x	.829	.831	.858
NetSMF	.939	.940	.949	.974	.977	.980	.987	.987	.987	.768	.786	.835
RandNE	.672	.700	.762	.833	.869	.903	.700	.739	.835	.507	.511	.514
LouvainNE	.820	.819	.815	.898	.899	.909	.735	.718	.746	.955	.954	.954
ProNE	.691	.761	.861	.623	.819	.908	.756	.803	.846	.847	.901	.944
Verse	**.957**	**.964**	**.971**	.880	.884	.858	**.988**	**.988**	**.988**	.951	.977	.988
HBDM	.899	.920	.935	.972	.979	.986	.897	.916	.932	.974	.980	.988
HBDM-RE	.940	.947	.957	**.980**	**.985**	**.988**	.962	.969	.971	**.976**	**.981**	**.989**

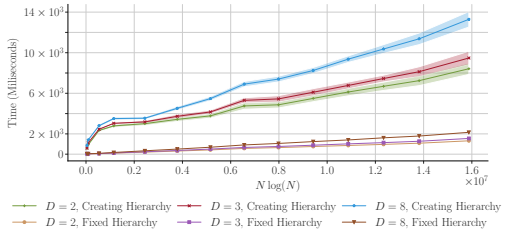


Fig. 5. *YouTube* network—HBDM-RE runtime complexity in miliseconds (ms) as a function of increasing sample sizes of network nodes (in terms of $N \log N$) until the sample set becomes the node set of the total graph. The y-axis showcases the average runtime over 100 iterations of the forward pass while the shaded areas provide standard deviations, as a measure of uncertainty. Runtimes are presented across $D = 2, 3, 8$ dimensions while we also show the runtime when the inferred hierarchy of the HBDM-RE is created from scratch versus when it is kept static.

model rapidly converges. After a couple of thousand iterations (very-scalable regime), we already obtain competitive performance for link prediction, which then gently increases until convergence. Our hyperparameter sensitivity analysis focuses on the predictive performance in the downstream task of link prediction. Since our method defines a likelihood over the network, the link predictive performance here shows how well the proposed framework characterizes generalizable patterns of network structure and we therefore focus the analysis on this aspect rather than node classification. If the network structure complies with the node classes, we can expect the node classification performance to follow the same behavior as the link prediction task. Potential disagreement in classification scores against the

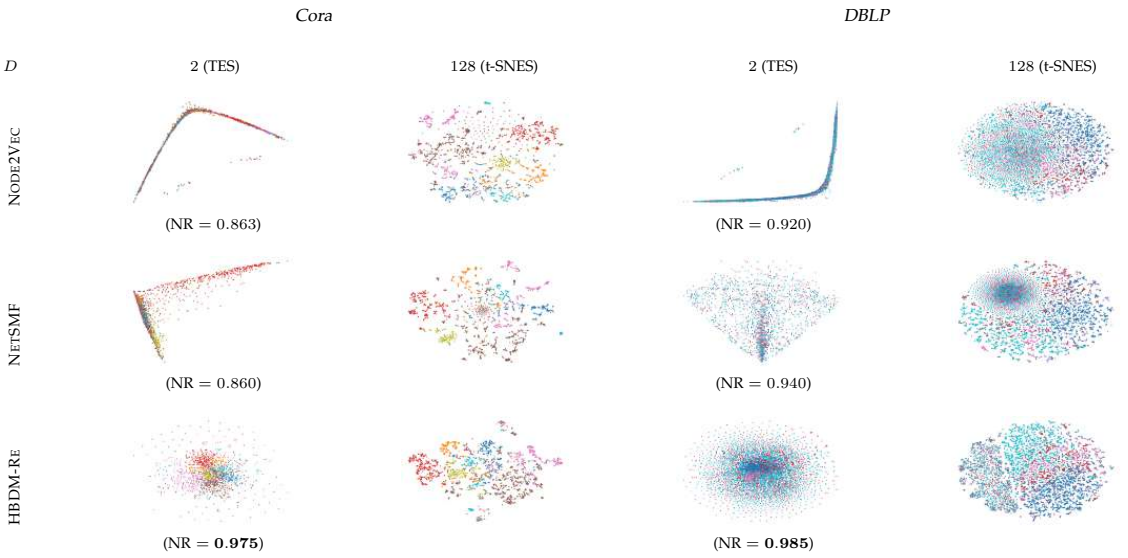


Fig. 6. 2-D true embedding space versus 128D t-SNE constructed space. For TES, we provide AUC-ROC for the network reconstruction (NR) task.

TABLE 6
Micro-F_1 scores varying dimension sizes for two moderate and large-scale networks.

	Cora			*DBLP*			*Amazon*			*YouTube*		
Dimension (D)	2	3	128	2	3	128	2	3	8	2	3	8
DEEPWALK	.502	.712	.838	.519	.605	.822	.231	.596	.929	.293	.351	.413
NODE2VEC	.419	.658	.835	.448	.540	.815	.096	.305	.895	-	-	-
LINE	.197	.191	.794	.328	.294	.771	.005	.003	.003	.185	.134	.177
NETMF	.389	.653	.835	.654	.707	.742	x	x	x	x	x	x
NETSMF	.554	.705	**.842**	.622	.732	**.829**	.387	.649	.845	.317	.361	.397
RANDNE	.271	.337	.731	.406	.473	.718	.223	.411	.787	.211	.226	.277
LOUVAINNE	.804	.811	.801	.780	.812	.825	**.970**	**.971**	**.974**	**.362**	.360	.359
PRONE	.450	.611	.830	.574	.634	.825	.420	.750	.933	.218	.274	.379
VERSE	.471	.719	.828	.518	.565	.757	.078	.416	.949	.243	.305	.393
LDM	**.810**	.802	.774	x	x	x	x	x	x	x	x	x
LDM-RE	.802	.803	.796	x	x	x	x	x	x	x	x	x
HBDM	.789	.807	.816	.812	.814	.772	**.970**	**.971**	.931	.320	.366	.414
HBDM-RE	.805	**.813**	.818	**.805**	**.822**	.808	.956	.955	.931	.326	**.367**	.414

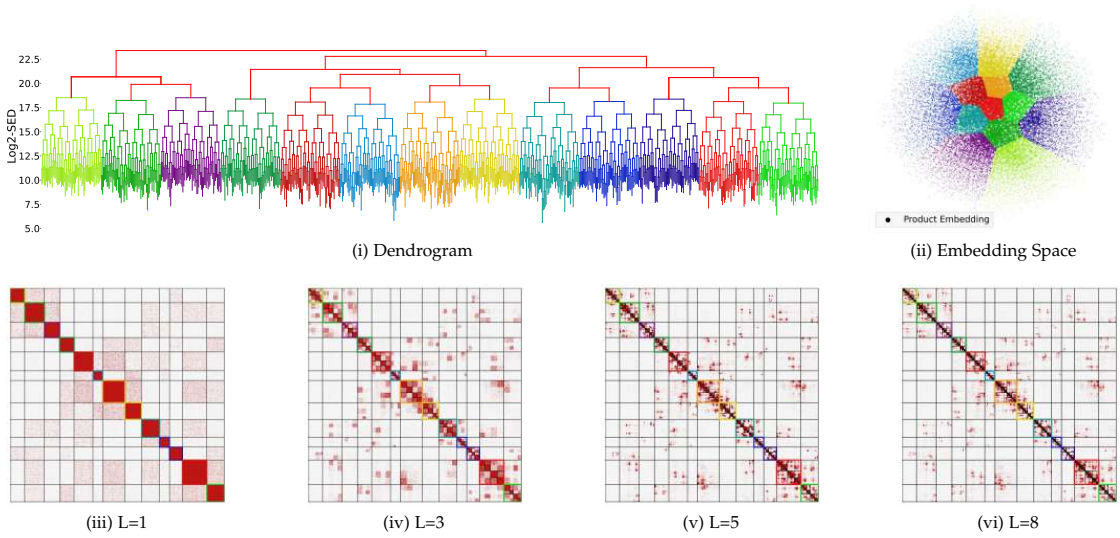
link prediction scores implies that the node labels do not follow the network structure, and such discrepancy would be network specific rather than a limitation of the method to be investigated.

3.3 Node classification

We assess the performance of the proposed framework in the uni/multi-label classification task and provide the Micro-F_1 scores in TABLE 6 (Macro-F_1 scores are reported in the supplementary). Scores are defined as the mean value over 10 random shuffles defining the training and test sets. Standard deviations as error bars were found in the scale of 10^{-3} and thus not presented. For the experimental setup, we randomly pick 50% of nodes as the training set and use the rest as the testing set. For an accurate comparison across different methods, we used two simple classifiers, a linear (logistic/multinomial regression classifier) and a non-linear (linearithmic k-nearest neighbors (*kNN*) classifier), and report the highest scores. We found that all methods benefit from using *kNN*. The number of neighbors was set to $k = 10$ (similar results were obtained with higher choices for k as well). Lastly, we report the average Micro-F_1 scores across 10 repeated trials. Results for the uni-labeled *Cora* and *DBLP* networks are reported in the two leftmost columns of TABLE 6. We observe that HBDM-RE and HBDM significantly outperform the baselines for the regimes of $D = 2, 3$ with only LOUVAINNE being competitive. Results for large-scale and multi-labeled networks *Amazon* and *YouTube* are provided by the two rightmost columns in TABLE 6. Again, the proposed framework outperforms the baselines for the low-dimensional regime with LOUVAINNE being on par. Comparing our framework with the classic LDM-RE and LDM, we again see an on-par performance which we attribute to the HBDM well preserving the intrinsic properties of homophily and transitivity. We further investigate the effect that the amount of training data has on classification performance. In Fig. 4 we provide the performance across multiple training size ratios and consider ultra-low dimensional embeddings of $D = 2, 3, 8$ for the *DBLP* network. We here observe that for the cases of $D = 2, 3$, our frameworks significantly outperform all the baselines with only LOUVAINNE being competitive. Increasing the dimensionality to $D = 8$, the baseline models are defined with enough capacity to be competitive while HBDM and HBDM-RE return favorable results.

3.4 Across tasks comparison:

We have considered multiple downstream tasks in each of which various baselines were found to be competitive against our HBDM frameworks. In general, the HBDM is characterized by the most consistent performance across tasks, especially for low dimensions. NETSMF was most competitive in the large-scale networks but underperformed



(i) Dendrogram (ii) Embedding Space

(iii) L=1 (iv) L=3 (v) L=5 (vi) L=8

Fig. 7. *Amazon* network dendrogram, embedding space and ordered adjacency matrices for the learned $D = 2$ embeddings of HBDM-RE and various levels (L) of the hierarchy.

TABLE 7
AUC scores for varying representation sizes over three bipartite networks.

	Drug-Gene			*GitHub*			*Gottron-Reuters*		
Dimension (D)	2	3	8	2	3	8	2	3	8
DEEPWALK	.673	.843	.878	.762	.853	.902	.673	.769	.905
NODE2VEC	.758	.814	.793	.724	.823	.876	.694	.766	.830
LINE	.798	.836	.867	.805	.766	.902	.715	.696	.850
NETMF	.576	.598	.742	.711	.711	.708	.747	.747	.730
NETSMF	.839	.838	.796	.846	.847	.857	.874	.934	.941
RANDNE	.536	.551	.613	.615	.651	.707	.769	.808	.872
LOUVAINNE	.760	.767	.779	.694	.702	.735	.654	.648	.679
PRONE	.667	.765	.831	.676	.771	.840	.606	.725	.909
VERSE	**.910**	**.913**	**.922**	**.943**	**.952**	**.959**	.962	.966	.967
HBDM	.798	.836	.889	.849	.869	.905	.941	.949	.950
HBDM-RE	.872	.891	.914	.932	.934	.937	**.964**	**.967**	**.973**

in the moderate-sized networks and node classification. VERSE was the most competitive baseline across tasks but massively underperformed in node classification for low dimensions. Furthermore, LOUVAINNE had high performance in node classification but underperformed in link prediction. Models such as NETSMF and VERSE can express structural (stochastic) equivalence while our HBDM explicitly expresses homophily and transitivity. This can explain the occasional higher performance of these baselines in the link prediction task.

3.5 Runtime complexity:

We assess the runtime complexity of the HBDM-RE framework in terms of increasing network sizes. In Fig. 5, we consider the *YouTube* network and show the runtime complexity in milliseconds (ms) as a function of increasing sample sizes of network nodes (in terms of $N \log N$) until the whole network is recovered. Runtimes are presented as the average across 100 iterations of the forward pass while the shaded areas provide the standard deviation. Runtimes are presented across $D = 2, 3, 8$ dimensions while we also compare the runtimes when the HBDM-RE builds the hierarchical structure via the k-means procedure and when the hierarchical structure is kept fixed. We here observe, that the runtime increases almost linearly as we increase the number of nertwork nodes. Comparing the runtime when creating the hierarchy via the k-means procedure from scratch versus keeping the dendrogram static from past iterations shows a significant decrease in runtime for the latter (in the experiments we create the hierarchy every 25'th iteration). Thus, the main bottleneck of the HBDM-RE approach is the computations required for the proposed Euclidean k-means procedure. Despite being deemed outside of the scope of this paper, such a bottleneck can be addressed by exploring existing procedures scaling conventional squared Euclidean k-means by avoiding unnecessary distance calculations [85] or by the use of binary space partitioning trees [86]. Such improved scaling would even admit utilization of non-binary splits beyond the root node improving the accuracy of the hierarchical approximation.

3.6 Network visualization

The graph representation learning literature mainly focuses on embeddings with dimensionality greater than $D = 2$ and 3. As a direct consequence, network visualizations rely on dimensionality reduction frameworks, typically using the t-distributed Stochastic Neighbor Embedding (t-SNE) [87]. In order to verify the quality of the t-SNE constructed Space (t-SNES), we provide the labeled-colored True Embedding Space (TES) in Fig. 6 for $D = 2$, as well as for $D = 128$ mapped to $D = 2$ via the use of t-SNE for *Cora* and *DBLP*. We see that the HBDM-RE frameworks

provide highly informative embeddings with no need for dimensionality reduction, unlike the rest of the baselines. This is also verified from the optimal performance in network reconstruction, HBDM-RE can successfully express the network structure using just $D = 2$. In Fig. 7 we provide the hierarchical block structure constructed by the HBDM-RE for the *Amazon* network. For visualization, we used the average within-cluster Euclidean distance to the centroid $\left(\Delta\{A,B\} = \frac{1}{N_A+N_B}\sum_{i\in C_A,C_B}\|\boldsymbol{z}_i - \boldsymbol{\mu}_{A\cup B}\|_2\right)$, as a linkage function to form a post-processing agglomerative clustering, for ordering the initial $\log N$ centroids. In Fig. 7 (i), we provide the dendrogram which denotes the agglomeration result in the top-level with red lines. The dendrogram continues with the hierarchical splits of our HBDM-RE where each color indicates the initial $\log N$ blocks. The y-axis of the dendrogram represents the binary logarithm of the Sum of Euclidean Distances, *Log2-SED* $= \log_2\left(\sum_{i\in C_k^{(l)}}\|\boldsymbol{z}_i - \boldsymbol{\mu}_k^{(l)}\|_2\right)$. Moreover, Fig. 7 (ii) conveys the corresponding latent space, colored based on the coarse $\log N$ split, revealing directly interpretable and accurate network representations. In Fig. 7 (iii), (iv), (v) and (vi) we showcase the organized adjacency matrices, based on the 2-dimensional HBDM-RE learned hierarchy for various levels L of the tree. We here, observe the representation power of the extracted hierarchy from just a 2-dimensional HBDM-RE defining communities and their sub-communities at finer and finer details.

For the bipartite case, we show how HBDM-RE can enhance our understanding of the bipartite structure at multiple scales and levels. Similar to the undirected case, Fig. 8 (i), indicates the dendrogram of the imposed hierarchy, enriched with agglomeration for a coarse level block ordering and proximity for the *GitHub* network. In addition, Fig. 8 (ii), provides the corresponding latent space, colored based on the coarse $\log N$ split. Notably, no dimensionality-reduction is necessary to define accurate network representation in the latent space of the two disjoint populations and visually access and express node similarity. In Fig. 8 (iii), (iv) and (v), we exhibit how the multi-scale structure evolves through different levels of the hierarchy defined by HBDM-RE, showcasing how a joint bi-clustering for complex network embeddings naturally can be obtained, with no need for post-processing steps. Our HBDM, can thus accurately characterize bipartite networks and successfully uncover their hierarchical block structure efficiently.

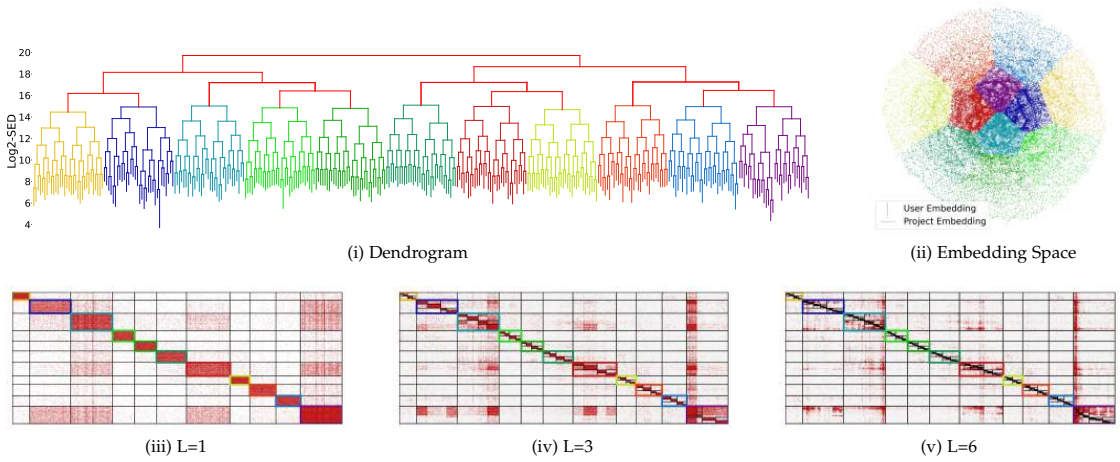
4 DISCUSSION

We developed the HBDM, a scalable reconciliation of latent distance models and their ability to account for homophily and transitivity with hierarchical representations of network structures. We demonstrated how the proposed HBDM provides favorable network representations by: (1) Operating with a Euclidean distance metric providing an intuitive human perception of node similarity. (2) Naturally representing multiscale hierarchical structure based on its block structure and carefully designed clustering procedure optimized in terms of Euclidean distances. (3) Directly and consistently operating in $D = 2, 3$ with high performance. (4) Performing well on all considered downstream tasks highlighting its ability to account for the underlying network structure. Importantly, the inferred hierarchical structure admits community discovery at multiple scales as highlighted by the inferred dendrograms and ordered adjacency matrices, and naturally extends to the characterization of communities of bipartite networks.

Our finding of ultra-low dimensional accurate characterizations of network structures supports the findings in [88] in which a logistic PCA model was found to enable exact low-dimensional recovery of multiple real-world networks. Whereas the work of [88] focuses on exact network reconstruction we find that generalizable patterns can be well extracted in ultra-low dimensional representations with performance saturating after just $D = 8$ dimensions for all networks considered. Whereas [88] found that their low-dimensional space did not perform well in classification tasks we observed strong node classification performance by the low-dimensional representations provided by HBDM. Importantly, for node classification, we observed better performance using KNN as opposed to simple linear classification based on logistic/multinomial regression typically used for node classification. This highlights that whereas most GRL works use linear classifiers there is no guarantee that the embedding space will be linearly separable and performance should therefore be compared to non-linear classifiers as they may provide more favorable performance as observed in this study.

Recent pioneering works [89], [90] have drawn significant attention of the research community by questioning the conventional embedding space preference. It is well known that many real-world networks show power-law degree distribution, or they can consist of latent hierarchical inner structures. Therefore, Euclidean space might not always be appropriate to represent such complex network architectures. It might also require higher-dimensional spaces to show comparable performance in the GRL tasks. The works of [89], [90] demonstrated that hyperbolic spaces, such as the Poincare disk model, can provide substantial benefits over the Euclidean space. The presented HBDM model naturally extends to other distance measures and future studies should explore how the HBDM can be extended to hierarchical representations beyond Euclidean geometry.

Covariate information plays an important role in the outstanding performance of GRL methods and especially GNNs. In the current LSM literature, side information is accounted for by extra regressors in the logit/log link functions expressing the likelihood of a dyad being connected. Using the Mahalanobis distance imposing a block-diagonal covariance matrix (see supplementary), the proposed HBDM can naturally incorporate covariate information directly to the latent space and notably construct multi-scale structures via the enriched and concatenated embedding of the latent variables and the covariate information. Our analysis presently did not explore side information and this is also why we did not include comparisons to prominent GNN-based approaches as these procedures do not provide favorable performance when only learning from the graph structure itself. As such, we observed (not shown) poor performance of GraphSage [91] when only having access to the graph structure in the present setup. The HBDM operates on static networks and thus is not naturally



(i) Dendrogram (ii) Embedding Space

(iii) L=1 (iv) L=3 (v) L=6

Fig. 8. *GitHub* network dendrogram, embedding space and ordered adjacency matrices for the learned $D = 2$ embeddings of HBDM-RE and various levels (L) of the hierarchy.

an inductive model. Nevertheless, potential new emerging nodes can be projected into the inferred latent space by fixing the embeddings of nodes present in the training set while optimizing the new nodes for their locations in the learned latent space. We leave a comparison of such a strategy against naturally inductive models such as GNNs for future work.

Our discoveries highlight the existence and importance of hierarchical multi-scale structures in complex networks. The across hierarchy re-ordered adjacency matrices given by HBDM, manifest sub-communities inside of what already appears as a strongly connected community. This points to how delicate the task of defining communities is and the importance of accounting for communities at multiple scales, as enabled by the HBDM. Importantly, these results generalize for bipartite networks where multi-scale geometric representations, joint hierarchical structures, and community discovery are arduous tasks.

The HBDM uses the LDM and thus is good at characterizing transitivity and homophily at a node and cluster level, whereas the random effects enable accounting for degree heterogeneity. Notably, the HBDM suffers from the limitations of the LDM and is thus unable to model stochastic equivalence. Future work should therefore investigate hierarchical structures imposed on more flexible GRL procedures enabling stochastic equivalence and contrast the performance when accounting for stochastic equivalence to the existing hierarchical methods based on the SBM [44], [45], [47]–[49], [52].

In conclusion, we proposed the Hierarchical Block Distance Model (HBDM), a scalable reconciliation of network embeddings using the latent distance model (LDM) and hierarchical characterizations of structure at multiple scales via a novel clustering framework. Notably, the model mimics the behavior of the LDM where the use of homophily and transitivity is most important while scaling in complexity by $\mathcal{O}(DN \log N)$. We analyzed thirteen networks from moderate sizes to large-scale with the HBDM having favorable performance when compared to existing scalable embedding procedures. In particular, we observed that the HBDM well predicts links and node classes providing accurate network visualizations and characterization of structure at multiple scales. Our results demonstrate that favorable performance can be achieved using ultra-low (i.e. $D = 2$) embedding dimensions and a scalable hierarchical representation that accounts for homophily and transitivity.

ACKNOWLEDGMENTS

We would like to express sincere appreciation and thank the reviewers for their constructive feedback and their insightful comments. We would also like to thank Louis Boucherie, Lasse Mohr Mikkelsen, and Giorgio Giannone for the valuable and fruitful discussions. The authors gratefully acknowledge the Independent Research Fund Denmark for supporting this work [grant number: 0136-00315B].

REFERENCES

[1] M. E. J. Newman, "The structure and function of complex networks," *SIAM Review*, vol. 45, no. 2, pp. 167–256, 2003.
[2] D. Liben-Nowell and J. Kleinberg, "The link prediction problem for social networks," in *CIKM*, 2003, p. 556–559.
[3] L. Getoor and B. Taskar, *Introduction to Statistical Relational Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2007.
[4] A. Grover and J. Leskovec, "Node2Vec: Scalable feature learning for networks," in *KDD*, 2016, pp. 855–864.
[5] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, no. 3, pp. 75–174, 2010.
[6] D. Zhang, J. Yin, X. Zhu, and C. Zhang, "Network representation learning: A survey," *IEEE Transactions on Big Data*, vol. 6, no. 1, pp. 3–28, 2020.
[7] B. Perozzi, R. Al-Rfou, and S. Skiena, "DeepWalk: Online learning of social representations," *CoRR*, vol. abs/1403.6652, 2014.
[8] J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, and Q. Mei, "LINE: Large-scale information network embedding," in *WWW*, 2015, pp. 1067–1077.
[9] A. Çelikkanat and F. D. Malliaros, "Exponential family graph embeddings," in *AAAI*, 2020, pp. 3357–3364.

[10] D. Nguyen and F. D. Malliaros, "BiasedWalk: Biased sampling for representation learning on graphs," in *Big Data*, 2018, pp. 4045–4053.

[11] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *NIPS*, 2013, pp. 3111–3119.

[12] S. Cao, W. Lu, and Q. Xu, "GraRep: Learning graph representations with global structural information," in *CIKM*, 2015, pp. 891–900.

[13] W. L. Hamilton, R. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *NIPS*, 2017.

[14] J. Zhang, Y. Dong, Y. Wang, J. Tang, and M. Ding, "Prone: Fast and scalable network representation learning," in *IJCAI*, 7 2019, pp. 4278–4284.

[15] J. Qiu, Y. Dong, H. Ma, J. Li, C. Wang, K. Wang, and J. Tang, "NetSMF: Large-scale network embedding as sparse matrix factorization," in *WWW*, 2019, pp. 1509–1520.

[16] A. K. Bhowmick, K. Meneni, M. Danisch, J.-L. Guillaume, and B. Mitra, "LouvainNE: Hierarchical louvain method for high quality and scalable network embedding," in *WSDM*, 2020, pp. 43–51.

[17] H. Chen, B. Perozzi, Y. Hu, and S. Skiena, "HARP: hierarchical representation learning for networks," in *AAAI*, 2018, pp. 2127–2134.

[18] Z. Zhang, P. Cui, H. Li, X. Wang, and W. Zhu, "Billion-scale network embedding with iterative random projection," in *ICDM*, 2018, pp. 787–796.

[19] Matias, Catherine and Robin, Stéphane, "Modeling heterogeneity in random graphs through latent space models: a selective review*," *ESAIM: Proc.*, vol. 47, pp. 55–74, 2014. [Online]. Available: https://doi.org/10.1051/proc/201447004

[20] D. K. Sewell and Y. Chen, "Latent space models for dynamic networks," *Journal of the American Statistical Association*, vol. 110, no. 512, pp. 1646–1657, 2015. [Online]. Available: https://doi.org/10.1080/01621459.2014.988214

[21] M. Salter-Townshend and T. H. McCormick, "Latent space models for multiview network data," *The Annals of Applied Statistics*, vol. 11, no. 3, pp. 1217 – 1244, 2017. [Online]. Available: https://doi.org/10.1214/16-AOAS955

[22] L. Zhu, D. Guo, J. Yin, G. V. Steeg, and A. Galstyan, "Scalable temporal latent space inference for link prediction in dynamic social networks," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 10, pp. 2765–2777, 2016.

[23] A. Çelikkanat, N. Nakis, and M. Mørup, "Piecewise-velocity model for learning continuous-time dynamic node representations," 2022. [Online]. Available: https://arxiv.org/abs/2212.12345

[24] P. Sarkar and A. W. Moore, "Dynamic social network analysis using latent space models," in *NIPS*, 2006, pp. 1145–1152.

[25] A. L. Smith, D. M. Asta, and C. A. Calder, "The Geometry of Continuous Latent Space Models for Network Data," *Statistical Science*, vol. 34, no. 3, pp. 428 – 453, 2019.

[26] J. Sosa and L. Buitrago, "A review of latent space models for social networks," *CoRR*, vol. abs/2012.02307, 2020.

[27] B. Kim, K. Lee, L. Xue, and X. Niu, "A review of dynamic network models with latent variables," 2017.

[28] N. Nakis, A. Çelikkanat, and M. Mørup, "Hm-ldm: A hybrid-membership latent distance model," 2022. [Online]. Available: https://arxiv.org/abs/2206.03463

[29] N. Nakis, A. Çelikkanat, L. Boucherie, C. Djurhuus, F. Burmester, D. M. Holmelund, M. Frolcová, and M. Mørup, "Characterizing polarization in social networks using the signed relational latent distance model," 2023. [Online]. Available: https://arxiv.org/abs/2301.09507

[30] P. D. Hoff, A. E. Raftery, and M. S. Handcock, "Latent space approaches to social network analysis," *JASA*, vol. 97, no. 460, pp. 1090–1098, 2002.

[31] H. Louch, "Personal network integration: transitivity and homophily in strong-tie relations," *Social Networks*, vol. 22, no. 1, pp. 45–64, 2000. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378873300000150

[32] C. Zhang, Y. Bu, Y. Ding, and J. Xu, "Understanding scientific collaboration: Homophily, transitivity, and preferential attachment," *J. Assoc. Inf. Sci. Technol.*, vol. 69, no. 1, p. 72–86, jan 2018. [Online]. Available: https://doi.org/10.1002/asi.23916

[33] P. BLOCK and T. GRUND, "Multidimensional homophily in friendship networks," *Network Science*, vol. 2, no. 2, p. 189–212, 2014.

[34] *Exponential Random Graph Models for Social Networks: Theory, Methods, and Applications*, ser. Structural Analysis in the Social Sciences. Cambridge University Press, 2012.

[35] P. N. Krivitsky, M. S. Handcock, A. E. Raftery, and P. D. Hoff, "Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models," *Social Networks*, vol. 31, no. 3, pp. 204 – 213, 2009.

[36] D. Cartwright and F. Harary, "Structural balance: a generalization of heider's theory." *Psychological review*, vol. 63 5, pp. 277–93, 1956.

[37] N. Friel, R. Rastelli, J. Wyse, and A. E. Raftery, "Interlocking directorates in irish companies using a latent space model for bipartite networks," *PNAS*, vol. 113, no. 24, pp. 6629–6634, 2016.

[38] M. Handcock, A. Raftery, and J. Tantrum, "Model-based clustering for social networks," *J R Stat Soc Ser A Stat Soc*, vol. 170, pp. 301 – 354, 03 2007.

[39] P. W. Holland, K. B. Laskey, and S. Leinhardt, "Stochastic blockmodels: First steps," *Social networks*, vol. 5, no. 2, pp. 109–137, 1983.

[40] K. Nowicki and T. A. B. Snijders, "Estimation and prediction for stochastic blockstructures," *JASA*, vol. 96, no. 455, pp. 1077–1087, 2001.

[41] P. D. Hoff, "Modeling homophily and stochastic equivalence in symmetric relational data," in *NIPS*, 2007, p. 657–664.

[42] A. Raftery, X. Niu, P. Hoff, and K. Y. Yeung, "Fast inference for the latent space network model using a case-control approximate likelihood," *J Comput Graph Stat*, vol. 21, 10 2012.

[43] E. Ravasz and A.-L. Barabási, "Hierarchical organization in complex networks," *Phys. Rev. E*, vol. 67, p. 026112, 2003.

[44] A. Clauset, C. Moore, and M. E. Newman, "Hierarchical structure and the prediction of missing links in networks," *Nature*, vol. 453, no. 7191, pp. 98–101, 2008.

[45] D. M. Roy, C. Kemp, V. Mansinghka, and J. Tenenbaum, "Learning annotated hierarchies from relational data," in *NIPS*, vol. 19, 2007.

[46] D. M. Roy and Y. Teh, "The mondrian process," in *NIPS*, D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, Eds., vol. 21. Curran Associates, Inc., 2008.

[47] T. Herlau, M. Mørup, M. N. Schmidt, and L. K. Hansen, "Detecting hierarchical structure in networks," in *CIP*. IEEE, 2012, pp. 1–6.

[48] T. Herlau, M. Mørup, and M. Schmidt, "Modeling temporal evolution and multiscale structure in networks," in *ICML*, 2013, pp. 960–968.

[49] T. P. Peixoto, "Hierarchical block structures and high-resolution model selection in large networks," *Physical Review X*, vol. 4, no. 1, 2014.

[50] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *J. Stat. Mech. Theory Exp.*, vol. 2008, no. 10, p. P10008, 2008.

[51] Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann, "Link communities reveal multiscale complexity in networks," *nature*, vol. 466, no. 7307, pp. 761–764, 2010.

[52] C. Blundell and Y. W. Teh, "Bayesian hierarchical community discovery," in *NIPS*, vol. 26, 2013.

[53] T. Li, L. Lei, S. Bhattacharyya, K. Van den Berge, P. Sarkar, P. J. Bickel, and E. Levina, "Hierarchical community detection by recursive partitioning," *JASA*, pp. 1–18, 2020.

[54] M. Wieling and J. Nerbonne, "Hierarchical spectral partitioning of bipartite graphs to cluster dialects and identify distinguishing features," in *TextGraphs-5*, 2010, pp. 33 – 41.

[55] I. S. Dhillon, "Co-clustering documents and words using bipartite spectral graph partitioning," in *KDD*, 2001, p. 269–274.

[56] Q. Cai and J. Liu, "Hierarchical clustering of bipartite networks based on multiobjective optimization," *IEEE Trans. Netw. Sci. Eng.*, vol. 7, no. 1, pp. 421–434, 2020.

[57] M. Rosvall and C. T. Bergstrom, "An information-theoretic framework for resolving community structure in complex networks," *PNAS*, vol. 104, no. 18, pp. 7327–7331, 2007.

[58] M. Mørup and M. N. Schmidt, "Bayesian community detection," *Neural computation*, vol. 24, no. 9, pp. 2434–2456, 2012.

[59] M. Handcock, A. Raftery, and J. Tantrum, "Model-based clustering for social networks," *J. R. Stat. Soc.*, vol. 170, pp. 301 – 354, 2007.

[60] P. D. Hoff, "Bilinear mixed-effects models for dyadic data," *JASA*, vol. 100, no. 469, pp. 286–295, 2005.

[61] D. K. Wind and M. Mørup, "Link prediction in weighted networks," in *2012 IEEE International Workshop on Machine Learning for Signal Processing*, 2012, pp. 1–6.

[62] B. Karrer and M. E. Newman, "Stochastic blockmodels and community structure in networks," *Physical review E*, vol. 83, no. 1, p. 016107, 2011.

[63] T. Herlau, M. N. Schmidt, and M. Mørup, "Infinite-degree-corrected stochastic block model," *Physical review E*, vol. 90, no. 3, p. 032819, 2014.
[64] A.-L. Barabási and M. Pósfai, *Network science*. Cambridge University Press, 2016.
[65] H. C. White, S. A. Boorman, and R. L. Breiger, "Social structure from multiple networks. i. blockmodels of roles and positions," *American journal of sociology*, vol. 81, no. 4, 1976.
[66] S. S. Epp, *Discrete Mathematics with Applications*, 4th ed. USA: Brooks/Cole, 2010.
[67] H. Tsutsu and Y. Morikawa, "An lp norm minimization using auxiliary function for compressed sensing," in *Proc. Int. Multiconf. Comp. Sci. Inf. Technol.*, 2012.
[68] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A K-Means clustering algorithm," *Applied Statistics*, vol. 28, no. 1, pp. 100–108, 1979.
[69] A. Tsitsulin, D. Mottin, P. Karras, and E. Müller, "VERSE," in *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*. ACM Press, 2018. [Online]. Available: https://doi.org/10.1145%2F3178876.3186120
[70] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017.
[71] P. Sen, G. Namata, M. Bilgic, L. Getoor, B. Gallagher, and T. Eliassi-Rad, "Collective classification in network data," *AI magazine*, 2008.
[72] J. Leskovec and J. J. Mcauley, "Learning to discover social circles in ego networks," in *NIPS*, 2012, pp. 539–547.
[73] J. Yang and J. Leskovec, "Defining and evaluating network communities based on ground-truth," *Knowledge and Information Systems*, vol. 42, no. 1, pp. 181–213, Jan 2015.
[74] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and analysis of online social networks," in *IMC*, 2007.
[75] R. Zafarani and H. Liu, "Social computing data repository at ASU," 2009.
[76] B. Perozzi, V. Kulkarni, H. Chen, and S. Skiena, "Don't walk, skip! online learning of multi-scale network embeddings," in *ASONAM*, 2017, pp. 258–265.
[77] J. Leskovec, J. Kleinberg, and C. Faloutsos, "Graph evolution: Densification and shrinking diameters," *ACM Trans. Knowl. Discov. Data*, vol. 1, no. 1, 2007.
[78] J. Leskovec and A. Krevl, "SNAP Datasets: Stanford large network dataset collection," Jun. 2014.
[79] J. Qiu, Y. Dong, H. Ma, J. Li, K. Wang, and J. Tang, "Network embedding as matrix factorization: Unifying DeepWalk, LINE, PTE, and Node2Vec," in *WSDM*, 2018.
[80] S. Dasgupta, "The hardness of k-means clustering," 2008.
[81] M. Mahajan, P. Nimbhorkar, and K. Varadarajan, "The planar k-means problem is np-hard," *Theoretical Computer Science*, vol. 442, pp. 13–21, 2012.
[82] D. S. Wishart, Y. D. Feunang, A. C. Guo, E. J. Lo, A. Marcu, J. R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, N. Assempour, I. Iynkkaran, Y. Liu, A. Maciejewski, N. Gale, A. Wilson, L. Chin, R. Cummings, D. Le, A. Pon, C. Knox, and M. Wilson, "DrugBank 5.0: a major update to the DrugBank database for 2018," *Nucleic Acids Research*, vol. 46, no. D1, pp. D1074–D1082, 11 2017.
[83] S. Chacon, "2009 github challenge," 2009. [Online]. Available: https://github.blog/2009-07-29-the-2009-github-contest/
[84] D. D. Lewis, Y. Yang, T. G. Rose, and F. Li, "Rcv1: A new benchmark collection for text categorization research," *J. Mach. Learn. Res.*, vol. 5, pp. 361–397, 2004.
[85] C. Elkan, "Using the triangle inequality to accelerate k-means," in *Proceedings of the 20th international conference on Machine Learning (ICML-03)*, 2003, pp. 147–153.
[86] D. Pettinger and G. Di Fatta, "Space partitioning for scalable k-means," in *2010 Ninth International Conference on Machine Learning and Applications*. IEEE, 2010, pp. 319–324.
[87] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008.
[88] S. Chanpuriya, C. Musco, K. Sotiropoulos, and C. E. Tsourakakis, "Node embeddings and exact low-rank representations of complex networks," *CoRR*, vol. abs/2006.05592, 2020.
[89] M. D. Ben Chamberlain and J. Clough, "Neural embeddings of graphs in hyperbolic space," in *MLG Workshop*, 2017.
[90] M. Nickel and D. Kiela, "Poincaré embeddings for learning hierarchical representations," in *NIPS*, vol. 30, 2017.
[91] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," *Advances in neural information processing systems*, vol. 30, 2017.

Nikolaos Nakis is currently a Ph.D. student at the Section for Cognitive Systems of the Technical University of Denmark. He received his BS in physics from the National and Kapodistrian University of Athens and his MS degree in mathematical modeling and computation from the Technical University of Denmark. His research mainly focuses on machine learning applied to complex systems and graph representation learning.

Abdulkadir Çelikkanat is currently a postdoctoral researcher at the Section for Cognitive Systems of the Technical University of Denmark. He completed his Ph.D. at the Centre for Visual Computing of CentraleSupélec, Paris-Saclay University, and he was also a member of the OPIS team at Inria Saclay. Before his Ph.D. studies, he received his Bachelor degree in Mathematics and Master's degree in Computer Engineering from Bogaziçi University. His research mainly focuses on the analysis of graph-structured data. In particular, he is interested in graph representation learning and its applications for social and biological networks.

Sune Lehmann Sune's work focuses on a quantitative understanding of social systems based on massive data sets. A physicist by training, his research draws on approaches from the physics of complex systems, machine learning, and statistical analysis. He works on large-scale behavioral data and while Sune's primary focus is on modeling complex networks, his research has made substantial contributions on topics such as human mobility, sleep, academic performance, complex contagion, epidemic spreading, and behavior on Twitter. He is the author of multiple high-impact papers and his research has won various prizes.

Morten Mørup received the MS and PhD degrees in applied mathematics from the Technical University of Denmark, where he is currently professor at the Section for Cognitive Systems at DTU Compute. He has been an associate editor of the IEEE Transactions on Signal Processing and his research interests include machine learning, neuroimaging, and complex network modeling.

HM-LDM: A Hybrid-Membership Latent Distance Model

Nikolaos Nakis, Abdulkadir Çelikkanat, and Morten Mørup

Section for Cognitive Systems,
Technical University of Denmark, Kongens Lyngby 2800, Denmark
nnak@dtu.dk, abce@dtu.dk, mmor@dtu.dk

Abstract. A central aim of modeling complex networks is to accurately embed networks in order to detect structures and predict link and node properties. The Latent Space Model (LSM) has become a prominent framework for embedding networks and includes the Latent Distance Model (LDM) and Eigenmodel (LEM) as the most widely used LSM specifications. For latent community detection, the embedding space in LDMs has been endowed with a clustering model whereas LEMs have been constrained to part-based non-negative matrix factorization (NMF) inspired representations promoting community discovery. We presently reconcile LSMs with latent community detection by constraining the LDM representation to the D-simplex forming the Hybrid-Membership Latent Distance Model (HM-LDM). We show that for sufficiently large simplex volumes this can be achieved without loss of expressive power whereas by extending the model to squared Euclidean distances, we recover the LEM formulation with constraints promoting part-based representations akin to NMF. Importantly, by systematically reducing the volume of the simplex, the model becomes unique and ultimately leads to hard assignments of nodes to simplex corners. We demonstrate experimentally how the proposed HM-LDM admits accurate node representations in regimes ensuring identifiability and valid community extraction. Importantly, HM-LDM naturally reconciles soft and hard community detection with network embeddings exploring a simple continuous optimization procedure on a volume constrained simplex that admits the systematic investigation of trade-offs between hard and mixed membership community detection.

Keywords: Latent Space Modeling, Community Detection, Non-negative Matrix Factorization, Graph Representation Learning.

1 Introduction

Networks naturally arise in the vast majority of scientific domains from physics to biology in order to model interactions among diverse entities with numerous instances such as collaboration, protein-protein, and brain connectivity networks [23]. Hence, graph analysis tools have become crucial to extract and analyze the underlying meaningful information from networks. In this direction, Graph

Representation Learning (GRL) [36] approaches have become a dominant way to carry out various downstream tasks such as node classification, link prediction, and community detection thanks to their superior performance compared to the classical techniques. GRL models mainly aim to map similar nodes in the network to close latent positions in a low dimension space by automatically learning corresponding node features [7].

The initial GRL works aimed to learn representations or features by simulating random walks over networks, taking inspiration from the Natural Language Processing field [4, 6, 25, 27, 31]. They mainly extract embeddings by optimizing the co-occurrence probability of node pairs within a certain distance through random walks. In recent years, we have witnessed a tremendous increase in the number of Graph Neural Networks (GNN) [7] methods with their usage in supervised tasks. They primarily rely on iterative message-passing operations of node attributes and hidden features around the surroundings of nodes for a given task. The matrix decomposition-based models [26, 27] are also a notable class of the GRL methods. They learn node embeddings by decomposing a designed target matrix based on first and higher-order proximity. However, few GRL methods rely on Non-negative Matrix Factorization (NMF), although it is a popular technique for unsupervised signal decomposition and approximation of multivariate non-negative data. NMF techniques have gathered lots of attention since they allow for structure retrieval through the latent factors of the imposed decomposition providing easy interpretable part-based representations [18].

Applications of NMF include network analysis allowing for efficient, unsupervised, and overlapping community detection, as well as GRL [2,20,33,34]. Within the NMF formulation, various works have sought to define mixed-membership frameworks for analysis and community detection purposes. A Mixed-Membership Stochastic Block Model (MM-SBM) [1] has been linked to the symmetric-NMF decomposition with uniqueness guarantees [20]. Standard least-squares NMF optimization was exchanged to a Poisson likelihood optimization for obtaining the propensity of nodes belonging to different communities [2]. In addition, a GRL approach for overlapping communities was presented in [33] where NMF was utilized to discover Poisson distributed mixed-memberships. These works, design mixed-memberships vectors for part-based representations [18] projected in an NMF constructed space where node similarity, as well as, position and metric properties, can be abstract.

The Latent Space Models (LSMs) are also one of the most powerful ways to learn latent representations [22]. These methods employ generalized linear models for constructing latent node embeddings which express important network characteristics. More specifically, the LDM [11] employs the Euclidean norm for positioning similar nodes closer in the latent space, which comes as a direct consequence of the triangular inequality, naturally representing transitivity (*"a friend of a friend is a friend"*) and homophily (*a tendency where similar nodes are more likely to connect to each other than dissimilar ones*) properties. The LDM can be generalized through the Eigenmodel [10] that can account for stochastic equivalence (*"groups of nodes defined by shared intra- and inter-group*

relationships") akin to the SBM [1] and the mixed membership SBM [1]. Furthermore, LDMs have been endowed with a clustering model imposing a Gaussian Mixture Model as prior forming the latent position clustering model [8, 29].

In this study, we propose a novel unsupervised representation learning method over graphs, namely, the Hybrid-Membership Latent Distance Model (HM-LDM), by bringing together the strengths of LDM and NMF. Specifically, the HM-LDM offers a reconciliation between part-based representations of networks and low-dimensional latent spaces satisfying similarity properties such as homophily and transitivity. The choice of these similarity properties is of high significance and one of the key characteristics behind GRL since they allow for easily interpretable discovery of network structure. Additionally, our proposed method permits us to capture the latent community structure of the networks using a simple continuous optimization procedure over the log-likelihood of the network. Notably, unlike most existing approaches imposing hard community memberships constraints, the assignment of community memberships in our proposed hybrid model can be controlled and altered through the simplex volume formed by the latent node representations. We extensively evaluate the performance of the proposed method in the ability to perform link prediction, as well as, community discovery over various networks of different types. We demonstrate that our model outperforms recent methods.
Source code: *Hybrid-Membership Latent Distance Model*.

2 Problem statement and proposed method

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be an undirected graph where $\mathcal{V}$ shows the vertex set and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ the edge set. We use $\mathbf{Y}_{N\times N} = (y_{i,j}) \in \{0,1\}^{N\times N}$ to denote the adjacency matrix of the graph where $y_{i,j} = 1$ if the pair $(i,j) \in \mathcal{E}$ otherwise it is equal to 0 for all $1 \leq i < j \leq N := |\mathcal{V}|$. Our main goal is to learn a representation, $\mathbf{w}_i \in \mathbb{R}^D$, for each node $i \in \mathcal{V}$ in a lower dimensional space ($D \ll N$) such that similar nodes in the network should have close embeddings. More specifically, we concentrate on mapping the nodes into the unit D-simplex set, $\Delta^D \subset \mathbb{R}_+^{D+1}$. Therefore, the extracted node embeddings can convey information about latent community memberships. Many GRL approaches also do not provide identifiable or unique solution guarantees, so their interpretation highly depends on the initialization of the hyper-parameters. In this study, we will also address the identifiability problem and seek identifiable solutions which can only be achieved up to a permutation invariance, as reported in Def. 1.

Definition 1 (Identifiabilty). *An embedding matrix* $\mathbf{W}$ *whose rows indicating the corresponding node representations is called an identifiable solution up to a permutation if it holds* $\widetilde{\mathbf{W}} = \mathbf{WP}$ *for a permutation* $\mathbf{P}$ *and a solution* $\widetilde{\mathbf{W}} \neq \mathbf{W}$.

We define a Poisson distribution over the adjacency matrix $\mathbf{Y}$ of the network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ to be conditionally independent given the unobserved latent positions,

and write the log-likelihood function as follows:

$$\log P(\mathbf{Y}|\boldsymbol{\Lambda}) = \sum_{\substack{i<j \\ y_{ij}=1}} \log(\lambda_{ij}) \; - \; \sum_{i<j} \Big(\lambda_{ij} + \log(y_{ij}!)\Big) \,. \tag{1}$$

where $\boldsymbol{\Lambda} = (\lambda_{ij})$ is the Poisson rate matrix which has absorbed the dependency over the model parameters. We here adopted a Poisson regression model similar to the work in [9]. In this study, we make use of a Poisson likelihood for modeling binary networks, as validated in [33].

We propose the Hybrid-Membership Latent Distance Model (HM-LDM) with a log-rate based on the ℓ^2-norm as:

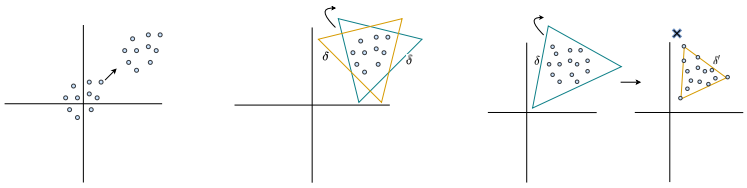
$$\log \lambda_{ij} = \Big(\gamma_i + \gamma_j - \delta^p \cdot ||\mathbf{w}_i - \mathbf{w}_j||_2^p\Big), \tag{2}$$

where $\mathbf{w_i} \in [0,1]^{D+1}$ and $\sum_{d=1}^{D+1} w_{id} = 1$, $\delta \in \mathbb{R}_+$ and $\gamma_i \in \mathbb{R}$ denotes the node-specific random-effects [9, 16] describing essentially the tendency of nodes to sending and receiving connections, accounting for degree heterogeneity. In addition, the norm degree $p \in \{1, 2\}$ controls the power of the ℓ^2-norm and combined with the latent embeddings sum-to-one condition constrains the latent space to the $D-$simplex with size equal to δ. A remarkable property of Eq. (2), for $p = 2$, is that it resembles a positive Eigenmodel with random effects: $\tilde{\gamma}_i + \tilde{\gamma}_j + (\tilde{\mathbf{w}}_i \boldsymbol{\Lambda} \tilde{\mathbf{w}}_j^\top)$ where $\boldsymbol{\Lambda}$ is a diagonal matrix having non-negative elements, i.e. $\tilde{\gamma}_i = \gamma_i - \delta^2 \cdot ||\mathbf{w}_i||_2^2$, $\tilde{\gamma}_j = \gamma_j - \delta^2 \cdot ||\mathbf{w}_j||_2^2$ and $\tilde{\mathbf{w}}_i \boldsymbol{\Lambda} \tilde{\mathbf{w}}_j^\top = 2\delta^2 \cdot \mathbf{w}_i \mathbf{w}_j^\top$ thus the squared Euclidean distance reconciles the conventional LDM and non-negativity constrained Eigenmodel. The squared Euclidean distance is not fully a metric but it still expresses homophily, leading to an interpretable latent space. Even though the triangle inequality is not exactly satisfied, it preserves the ordering of pairwise Euclidean distances, and it is highly preferred in applications since it is a strictly convex smooth function. By the well-known cosine formula, we have

$$||\mathbf{w}_i - \mathbf{w}_j||_2^2 = ||\mathbf{w}_i - \mathbf{w}_k||_2^2 + ||\mathbf{w}_k - \mathbf{w}_j||_2^2 - 2||\mathbf{w}_i - \mathbf{w}_k||_2 ||\mathbf{w}_k - \mathbf{w}_j||_2 \cos(\theta),$$

where $\theta \in (-\pi/2, \pi/2)$ is the angle between $\mathbf{w}_i - \mathbf{w}_k$ and $\mathbf{w}_k - \mathbf{w}_j$. Notice that the third term also converges to 0 for $\theta \to \pi/2$. For the case where $\theta \in [-\pi/2, \pi/2]$, it holds the triangle inequality: $||\mathbf{w}_i - \mathbf{w}_j||_2^2 \leq ||\mathbf{w}_i - \mathbf{w}_k||_2^2 + ||\mathbf{w}_k - \mathbf{w}_j||_2^2$.

The embedding vectors, $\{\mathbf{w}_i\}_{i=1}^N$ in Eq. (2), are constrained to non-negative values and to sum to one. Thereby, they reside on a simplex showing the participation of node $i \in \mathcal{V}$ over $D + 1$ latent communities. Any LDM can be translated to the non-negative orthant without any loss in performance or in expressive capability. Non-negative embeddings do not affect the distance metric, as it is invariant to translation, as shown by Figure 1 (a). In addition, the D-dimensional non-negative orthant can be reconstructed by a large enough D-simplex. Based on these arguments, it is trivial to show that for large values of the δ parameter in Eq. (2), despite the sum-to-one constraint on the embeddings $\mathbf{W}$, we obtain an unconstrained LDM, as distances are unbounded when



(a) Translation invariances.

(b) Rotation invariances.

(c) Decreased simplex volume ensuring identifiability.

Fig. 1: A 2-dimensional latent space with the 2-simplex given as the green and yellow triangles, the blue points denote embedding positions of the LDM and δ is the simplex size.

$\delta \to +\infty$. In this case, the memberships defined by $\mathbf{W}$ are not uniquely identifiable due to the distance invariance of rotation, as seen in Figure 1 (b). However, by shrinking the simplex volume (equivalent to decreasing δ), eventually, the D-dimensional space of LDM will no longer be enclosed inside the D-simplex, forcing nodes to start populating the corners of this smaller simplex. We call a node *champion* if its latent representation is a standard binary unit vector.

Definition 2 (Community champion). *A node for a latent community is called champion if it belongs to the community (simplex corner) while forming a binary unit vector.*

The champion nodes are of great significance for identifiability because if every corner of the simplex is populated by at least one node (champion), then the solution of the model is identifiable (up to a permutation matrix) (Def. 1) as any random rotation does not leave the solution invariant anymore, as shown by Figure 1 (c). We observe then, that the scalar, δ, controls the type of memberships of the model and its expressive capabilities. Large enough values lead to the basic LDM but inherits its rotational invariance. Small values of δ lead to identifiable solutions and ultimately hard cluster assignments. Thereby, for very small values of δ, nodes are solely assigned to the simplex corners. Lastly, we can also find regimes of values for δ that offer identifiable solutions but also performance similar to LDM, defining a silver lining.

A different take on the identifiability of the model for $p = 2$, can also be given under the Non-negative Matrix Factorization (NMF) theory. This is easily shown by a re-parameterization of Eq. (2) by $\tilde{\gamma}_i + \tilde{\gamma}_j + 2\delta^2 \cdot (\mathbf{w}_i \mathbf{w}_j^\top)$ as described in Eq. (2). In this formulation, the product $\mathbf{W}\mathbf{W}^\top$ defines a symmetric NMF problem which is an identifiable and unique factorization (up to permutation invariance) when $\mathbf{W}$ is full-rank and at least one node resides solely in each simplex corner, ensuring separability [12,20]. Under this NMF formulation, the product $\mathbf{w}_i \mathbf{w}_j^\top \in [0, 1]$ achieves its upper bound only if both nodes i and j reside in the same corner of the simplex. The parameter, δ, acts as a simple multiplicative factor in the first

Table 1: Network statistics; $|\mathcal{V}|$: # Nodes, $|\mathcal{E}|$: # Edges, $|\mathcal{K}|$: # Communities.

	AstroPh [19]	*GrQc* [19]	*Facebook* [19]	*HepTh* [19]	*Hamilton* [21]	*Amherst* [21]	*Rochester* [21]	*Mich* [21]
$\lvert\mathcal{V}\rvert$	17,903	5,242	4,039	8,638	2,118	2,021	4,145	2,933
$\lvert\mathcal{E}\rvert$	197,031	14,496	88,234	24,827	87,486	87,496	145,305	54,903
$\lvert\mathcal{K}\rvert$	-	-	-	-	15	15	19	13

term of the objective function of HM-LDM, given in Eq. (1), while in the second term acts as a power of the exponential function. For small values of δ, the model is biased towards hard latent community assignments of nodes since similar nodes achieve high rates only when they belong to the same latent community (simplex corner). On the other hand, nodes heading towards the simplex corners for large values of δ lead to an exponential change in the second term of the log-likelihood function given in Eq. (1). Thus, a possible hard allocation of dissimilar nodes to the same community penalizes the likelihood severely. For this reason, high order of δ benefits mixed-membership allocations.

3 Experimental evaluation

We proceed by evaluating the efficiency and performance of the proposed method. In our set-up, we make use of networks with unknown community structures, as well as, with ground-truth communities. We employ the former networks to validate the ability of our framework to discover identifiable latent structures and predict missing links. The latter networks are used to verify that the HM-LDM discovers communities successfully. We consider multiple social and scientific collaboration networks as shown by Table 1. We treat all networks as unweighted and undirected.

For the training of HM-LDM we optimize the log-likelihood function of Eq. (1) via the Adam optimizer [15] with learning rate $lr \in [0.01, 0.1]$. The node-specific random effects vector $\boldsymbol{\gamma} \in \mathbb{R}^N$ is randomly initialized and then tuned alone by optimizing a Poisson log-likelihood with a rate as $\log \lambda_{ij} = \gamma_i + \gamma_j$. Next, the latent embeddings matrix $\mathbf{W}$ is initialized based on the eigenvalues obtained by the spectral decomposition of the normalized Laplacian matrix of the network [13, 24]. In all experiments, we compare against unsupervised methods, and we do not include GNNs since they perform poorly in unsupervised tasks due to the over-smoothing effect [35].

Link prediction: For the link prediction experiments, we follow the well-established strategy [6, 25] and remove 50% of the network edges while keeping the residual network connected. The removed edges combined with a sample of the same number of node pairs (which are not the edges of the original network) construct the negative instances for the testing set. We utilize the residual network to learn the node embeddings.

We consider four networks with unknown community structures and asses performance across different dimensions. In Table 2, we compare the results of our method with other prominent GRL and NMF approaches in terms of the

Table 2: Area Under Curve (AUC-ROC) scores for varying representation sizes.

	AstroPh			*GrQc*			*Facebook*			*HepTh*		
Dimension (D)	8	16	32	8	16	32	8	16	32	8	16	32
DeepWalk [25]	.945	.950	.952	.919	.916	.929	.986	.986	.984	.874	.867	.873
Node2Vec [6]	.950	.962	.957	.897	.913	.930	.988	.988	.987	.881	.882	.881
LINE [31]	.909	.938	.947	.920	.925	.919	.981	.987	.983	.873	.886	.882
NetMF [27]	.813	.823	.839	.860	.866	.877	.935	.963	.971	.792	.806	.821
NetSMF [26]	.891	.901	.919	.837	.858	.886	.975	.981	.985	.809	.822	.836
LouvainNE [3]	.813	.811	.819	.868	.875	.873	.958	.961	.963	.874	.867	.873
ProNE [37]	.907	.929	.947	.885	.911	.921	.971	.982	.987	.827	.846	.859
NNSED [30]	.861	.882	.891	.792	.808	.828	.908	.927	.935	.756	.779	.796
MNMF [32]	.893	.925	.943	.911	.928	.937	.965	.978	.982	.857	.880	.891
BigClam [34]	.500	.723	.810	.752	.769	.780	.744	.722	.647	.776	.700	.748
SymmNMF [17]	.767	.779	.800	.729	.772	.835	.933	.942	.951	.696	.727	.766
HM-LDM($p=1$)	.956	.952	.952	**.944**	**.948**	**.951**	.982	.979	.974	**.916**	**.921**	**.924**
HM-LDM($p=2$)	**.972**	**.973**	**.963**	.940	.942	.946	**.992**	**.993**	**.993**	.908	.910	.911

Area Under Curve-Receiver Operating Characteristic (AUC-ROC). All baselines have been tuned and feature vectors for dyads are constructed based on binary operators (average, Hadamard, weighted-L1, weighted-L2) [6]. For these constructed feature vectors we further train a logistic regression model with L_2 regularization to make predictions. In particular, for the baselines we choose the hyperparameter settings for each model, as well as, the binary operator for which the logistic regression predictions return the maximum AUC-ROC score.

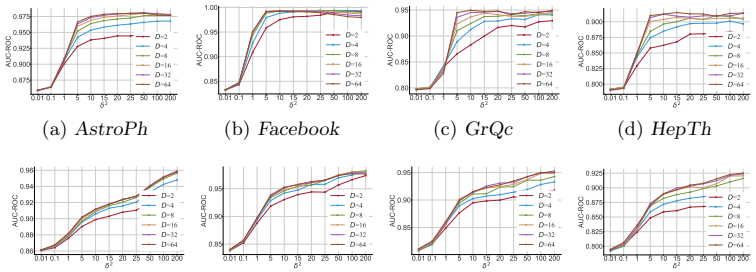
In contrast, for our models, we adopt an unbiased evaluation, and we choose the first of the considered δ values which keeps the solution identifiable (at least one champion per community), as δ decreases. We note though, the existence of identifiable regimes with higher predictive power. Furthermore, predictions and AUC-ROC scores for HM-LDM, can be obtained directly (without the use of a logistic regression model) and are based on the learned Poisson rates λ_{ij} of the test set pairs $\{i, j\}$. The true dimensions for HM-LDM are $D + 1$ but reported as D since they express the true number of model parameters, for a fair comparison with the baselines. For our method, we show the mean performance over five independent runs (error bars were found to be in the scale of 10^{-3} and thus not presented).

Comparing now the results with the non-NMF models, we observe that our HM-LDM (either $p = 1$ or $p = 2$) outperforms the baselines and in most cases significantly, returning favorable results. For the NMF models, we see mostly a big performance gap with the HM-LDM, showcasing the existence of regimes for δ where we can successfully achieve identifiable community memberships while also exhibiting the link prediction power of the LDM. (AUC Precision-Recall scores are similar to the AUC-ROC scores and thus not presented)

Performance and simplex sizes: In Figure 2, we provide the link prediction performance as a function of δ^2 in terms of the AUC-ROC scores across various latent dimensions, networks and for both $p = 1$ and $p = 2$. We here

observe that small δ values provide the minimum scores. This phenomenon is anticipated due to the fact that homophily properties are not sufficiently met (except within clusters) due to the very small simplex volume that these low δ values define. Rethinking HM-LDM with $p = 2$ as a positive Eigenmodel, we can also notice how the positivity constraint on the Λ diagonal matrix does not allow for stochastic equivalence properties which would essentially boost performance even on low simplex volumes. As we increase the values of δ, we naturally reach the performance of an unconstrained LDM. Comparing now, the squared and simple ℓ^2-norm metric we observe that the former converges to performance saturation more rapidly.

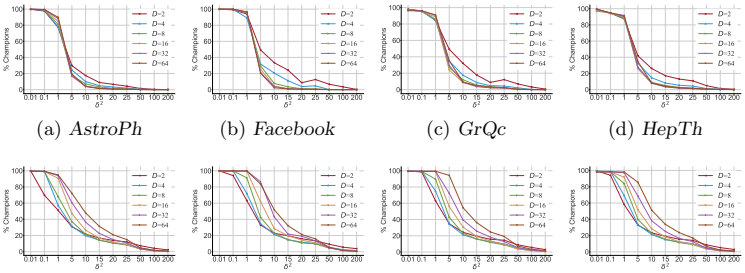
Type and quality of latent memberships: In order to understand how the size of the simplex affects the membership types of HM-LDM, we provide in Figure 3 the total network percentage of community champions as a function of δ^2 across various latent dimensions. As expected, for very small values of δ almost 100% of nodes are assigned solely to a unique simplex corner, yielding hard cluster assignments. As we increase δ, we observe that more and more nodes are assigned with mixed-memberships; on the other hand, the number of champions goes to zero across all dimensions for large values of δ. Contrasting again, the different powers p of the HM-LDM formulation, we notice that the decrease in community champions is steeper for $p = 2$. This also explains why the squared ℓ^2 choice leads to faster convergence in the AUC-ROC, as the model converges faster to the classic LDM. Overall, it is evident that the $p = 2$ HM-LDM needs smaller simplex volumes to be identifiable. We continue with assessing unique latent structures of HM-LDM. For that purpose, in Figure 4 we provide the reorganized adjacency matrices with respect to the community allocations of HM-LDM (for mixed-memberships we assign a node based on the maximum membership). We witness how HM-LDM successfully discovers latent commu-



(a) *AstroPh* (b) *Facebook* (c) *GrQc* (d) *HepTh*

(e) *AstroPh* (f) *Facebook* (g) *GrQc* (h) *HepTh*

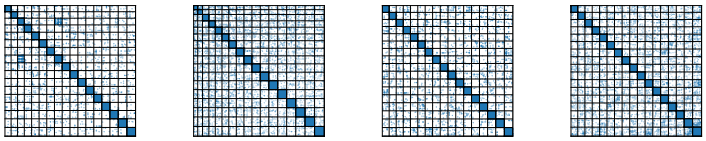
Fig. 2: AUC-ROC scores as a function of δ^2 across dimensions for HM-LDM. Top row: $p = 2$. Bottom row $p = 1$.



(a) *AstroPh* (b) *Facebook* (c) *GrQc* (d) *HepTh*

(e) *AstroPh* (f) *Facebook* (g) *GrQc* (h) *HepTh*

Fig. 3: Total community champions (%) in terms of δ^2 across dimensions for HM-LDM. Top row: $p = 2$. Bottom row $p = 1$.



(a) *GrQc* ($p = 2$) (b) *HepTh* ($p = 2$) (c) *GrQc* ($p = 1$) (d) *HepTh* ($p = 1$)

Fig. 4: Ordered adjacency matrices based on the memberships of a $D = 16$ dimensional HM-LDM with δ values ensuring identifiability.

nities, facilitating part-based network representations while choosing appropriate δ regimes ensure identifiability.

Experiments using real ground-truth communities: In order to assess the ability of HM-LDM to discover informative communities, we make use of four networks providing ground-truth community labels. For the NMF-based methods, including ours, we test the ability of the algorithms to detect valid structures by comparing the inferred memberships with the ground-truth community labels while we set the latent dimensions to be equal to the total number of communities. For the GRL approaches which do not define memberships, we extract latent embeddings and use *k-means* (average over 20 runs for robustness) to obtain memberships. We report the Normalized Mutual Information (NMI) score, as well as, the Adjusted Rand Index (ARI), both measures have been validated for community quality assessment in [5]. Again, all the baselines have been tuned individually for each network in terms of their hyperparameters. In contrast, for our HM-LDM, we do not perform any tuning and we just set $\delta = 1$ for all networks since this choice provides in general informative and mostly hard

Table 3: Normalized Mutual Information (NMI) and Adjusted Rand Index (ARI) scores for networks with ground-truth communities.

	Amherst		*Rochester*		*Mich*		*Hamilton*	
Metric	NMI	ARI	NMI	ARI	NMI	ARI	NMI	ARI
DeepWalk [25]	.498	.347	.348	.205	.207	.157	.447	.303
Node2Vec [6]	.535	.375	.364	.223	.217	.161	.481	.348
LINE [31]	.549	.452	.365	.217	**.249**	.192	.499	.411
NetMF [27]	.491	.330	.377	.243	.237	.136	.456	.297
NetSMF [26]	.562	.408	.381	.228	.242	.169	.494	.391
LouvainNE [3]	.562	.395	.347	.204	.175	.114	.475	.334
ProNE [37]	.536	.443	.356	.312	.229	.200	.478	.396
NNSED [30]	.295	.243	.168	.116	.064	.035	.335	.285
MNMF [32]	.542	.362	.324	.171	.188	.102	.466	.287
BigClam [34]	.091	.066	.028	.022	.024	.015	.053	.041
SymmNMF [17]	**.596**	.397	.308	.175	.207	.088	.437	.341
HM-LDM($p=1$)	.562	.502	**.400**	**.392**	.228	**.205**	**.527**	.485
HM-LDM($p=2$)	.539	**.506**	.384	.373	.217	.183	.507	**.504**

cluster assignments. For our method and the classic LDMs, we report scores averaged over five independent runs in each of which we run the algorithm five times extracting the model with the lowest training loss to remove the effect of local-minimas. We summarize our findings in Table 3, where we witness mostly favorable or on-par performance of HM-LDM with all of the competitive baselines for the NMI metric. For the ARI metric, we observe that our framework outperforms significantly the baselines in all of the considered networks.

Comparison with the LDM: We further investigate the performance of HM-LDM against the LDM, including random effects for a fair comparison and for both normal and squared ℓ^2-norm LDM-Re and LDM-Re-$(\ell^2)^2$, respectively. Towards that aim, in Table 4 and Table 5 we provide the performance scores for the link prediction and clustering tasks of each model. We here witness that constraining the latent space in identifiable simplex volumes leads to a minor decrease in the predictive power, in terms of the AUC-ROC. For the community detection task, we see favorable NMI scores while the HM-LDM leads to considerably higher ARI scores. Comparing the classical LDM with HM-LDM for $\delta^2 = 10^3$ provides on-par link-prediction performance but the clustering scores drop significantly. This is expected as for large simplex volumes the HM-LDM approximates almost exactly the LDM with the cost of identifiability.

Extension to bipartite networks: Finally, we showcase the extension of our HM-LDM framework to the analysis of bipartite networks. This is straightforward by introducing a different set of latent variables for the two disjoint sets of nodes, as defined by the bipartite structure. In particular, HM-LDM for $p = 2$, simply extends the symmetric NMF formulation, obtained for the undirected networks, to the non-symmetric NMF specification. In Figure 5, we provide the re-ordered adjacency matrix with respect to the community allocations defined by the learned embeddings of HM-LDM for a *Drug-Gene* [19]

Table 4: HM-LDM and LDM-RE comparison for the link prediction task.

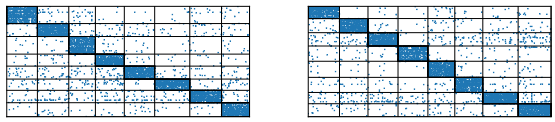
	AstroPh			*GrQc*			*Facebook*			*HepTh*		
Dimension (D)	8	16	32	8	16	32	8	16	32	8	16	32
LDM-RE	.973	.974	.979	.949	.952	.954	.993	.994	.992	.920	.923	.923
HM-LDM($p=1, \delta^2=$ identifiable)	.956	.952	.952	.944	.948	.951	.982	.979	.974	.916	.921	.924
HM-LDM($p=1, \delta^2=10^3$)	.967	.967	.965	.956	.955	.951	.985	.986	.987	.932	.931	.926
LDM-RE-$(\ell^2)^2$	.979	.978	.976	.944	.944	.945	.990	.990	.991	.913	.912	.909
HM-LDM($p=2, \delta^2=$ identifiable)	.972	.973	.963	.940	.942	.946	.992	.993	.993	.908	.910	.911
HM-LDM($p=2, \delta^2=10^3$)	.984	.983	.980	.948	.946	.946	.991	.991	.992	.920	.918	.913

Table 5: HM-LDM and LDM-RE comparison for the clustering task.

	Amherst		*Rochester*		*Mich*		*Hamilton*	
Metric	NMI	ARI	NMI	ARI	NMI	ARI	NMI	ARI
LDM-RE	.548	.366	.391	.212	.230	.132	.491	.320
HM-LDM($p=1, \delta^2=$ identifiable)	.562	.502	.400	.392	.228	.205	.527	.485
HM-LDM($p=1, \delta^2=10^3$)	.439	.386	.308	.303	.176	.133	.405	.377
LDM-RE-$(\ell^2)^2$	.546	.370	.393	.211	.231	.137	.497	.327
HM-LDM($p=2, \delta^2=$ identifiable)	.539	.506	.384	.373	.217	.183	.507	.504
HM-LDM($p=2, \delta^2=10^3$)	.240	.133	.206	.119	.116	.056	.232	.209

network ($|\mathcal{V}| = 7,341|$, $|\mathcal{E}| = 15,138$) where we observe a clear block structure. Importantly, the HM-LDM offers identifiable joint embedding representations, mixed memberships, and community discovery for bipartite networks, tasks considered to be non-trivial and arduous.

Complexity analysis: The HM-LDM framework requires the computation of the node pairwise distance matrix and consequently scales prohibitively as $\mathcal{O}(N^2)$ in time and space. Fortunately, there are various ways of scaling HM-LDM for the analysis of large-scale networks. One way is through unbiased estimators of the log-likelihood given by Eq. (1). This is possible through random sampling a set of network nodes S (per iteration) and taking a gradient step based on the log-likelihood of the block defined by the sampled node-set, returning an $\mathcal{O}(S^2)$ space and time complexity. Another option is through the case-control approach [28] scaling on the number of network edges as $\mathcal{O}(E)$.



(a) $p = 1$, $\delta = 1$ (b) $p = 2$, $\delta = 1$

Fig. 5: *Drug-Gene* ordered adjacency matrices based on HM-LDM with $D = 8$.

Lastly, the Hierarchical Block Distance Model (HBDM) [22] is an attractive option where gradient steps over the model parameters are based on a hierarchical approximation of the likelihood of the whole network. The HBDM model scales linearithmicly as $\mathcal{O}(N \log N)$ both in space and time while also offering hierarchical characterizations of structures at multiple scales.

4 Conclusion and future work

In this paper, we have proposed the HM-LDM that reconciles network embedding and latent community detection. The approach utilizes both the normal and squared Euclidean distance model where the latter integrated the non-negativity constrained Eigenmodel with the Latent Distance Model. We demonstrated that the model could be constrained to the simplex without losing expressive power. The reduced simplex provides unique representations, ultimately resulting in hard clustering of nodes to communities when the simplex is sufficiently shrunk. Notably, the proposed HM-LDM combines network homophily and transitivity properties with latent community detection enabling explicit control of soft and hard assignment through the volume of the induced simplex. We observed favorable link prediction performance in regimes in which the HM-LDM provides unique representations while enabling the ordering of the adjacency matrix in terms of prominent latent communities. Finally, we showed the ability of the model to extract correct community structures across multiple networks and showcased how the analysis extends to bipartite networks. Future work should compare the performance of HM-LDM against classical non-embedding methods such as the Degree Corrected Stochastic Block Model (DC-SBM) [14] or the Mixed Membership Stochastic Block Model (MM-SBM) [1]. Such a comparison is of particular interest since DC-SBM accounts for degree heterogeneity while MM-SBM for soft assignments, two important properties of HM-LDM.

Acknowledgements

We would like to thank the reviewers for the constructive feedback and their insightful comments. We would also like to thank Sune Lehmann, Louis Boucherie, Lasse Mohr Mikkelsen, and Giorgio Giannone for the useful and fruitful discussions. We gratefully acknowledge the Independent Research Fund Denmark for supporting this work [grant number: 0136-00315B].

References

1. Airoldi, E.M., Blei, D.M., Fienberg, S.E., Xing, E.P.: Mixed membership stochastic blockmodels. J Mach Learn Res 9(65), 1981–2014 (2008)
2. Ball, B., Karrer, B., Newman, M.E.J.: An efficient and principled method for detecting communities in networks. CoRR abs/1104.3590 (2011)

3. Bhowmick, A.K., Meneni, K., Danisch, M., Guillaume, J.L., Mitra, B.: LouvainNE: Hierarchical louvain method for high quality and scalable network embedding. In: WSDM. pp. 43–51 (2020)
4. Çelikkanat, A., Malliaros, F.D.: Exponential family graph embeddings. In: AAAI. pp. 3357–3364 (2020)
5. Chakraborty, T., Dalmia, A., Mukherjee, A., Ganguly, N.: Metrics for community analysis: A survey (2016)
6. Grover, A., Leskovec, J.: Node2Vec: Scalable feature learning for networks. In: KDD. pp. 855–864 (2016)
7. Hamilton, W.L., Ying, R., Leskovec, J.: Representation learning on graphs: Methods and applications. IEEE Data Eng. Bull. 40(3), 52–74 (2017)
8. Handcock, M.S., Raftery, A.E., Tantrum, J.M.: Model-based clustering for social networks. J R Stat Soc Ser A Stat Soc. 170(2), 301–354 (2007)
9. Hoff, P.D.: Bilinear mixed-effects models for dyadic data. JASA 100(469), 286–295 (2005)
10. Hoff, P.D.: Modeling homophily and stochastic equivalence in symmetric relational data (2007)
11. Hoff, P.D., Raftery, A.E., Handcock, M.S.: Latent space approaches to social network analysis. JASA 97(460), 1090–1098 (2002)
12. Huang, K., Sidiropoulos, N.D., Swami, A.: Non-negative matrix factorization revisited: Uniqueness and algorithm for symmetric decomposition. IEEE Trans. Signal Process 62(1), 211–224 (2014)
13. Jianbo Shi, Malik, J.: Normalized cuts and image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(8), 888–905 (2000)
14. Karrer, B., Newman, M.E.: Stochastic blockmodels and community structure in networks. Physical review E 83(1), 016107 (2011)
15. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization (2017)
16. Krivitsky, P.N., Handcock, M.S., Raftery, A.E., Hoff, P.D.: Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models. Social Networks 31(3), 204 – 213 (2009)
17. Kuang, D., Ding, C., Park, H.: Symmetric nonnegative matrix factorization for graph clustering. In: SDM (2012)
18. Lee, D.D., Seung, H.S.: Learning the parts of objects by nonnegative matrix factorization. Nature 401, 788–791 (1999)
19. Leskovec, J., Krevl, A.: SNAP Datasets: Stanford large network dataset collection (Jun 2014)
20. Mao, X., Sarkar, P., Chakrabarti, D.: On mixed memberships and symmetric nonnegative matrix factorizations. In: ICML. vol. 70 (2017)
21. Mucha, P., Porter, M.: Social structure of facebook networks. Physica A: Statistical Mechanics and its Applications 391, 4165–4180 (08 2012)
22. Nakis, N., Çelikkanat, A., Jørgensen, S.L., Mørup, M.: A hierarchical block distance model for ultra low-dimensional graph representations (2022)
23. Newman, M.E.J.: The structure and function of complex networks. SIAM Review 45(2), 167–256 (2003)
24. Ng, A.Y., Jordan, M.I., Weiss, Y.: On spectral clustering: Analysis and an algorithm. In: Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic. p. 849–856. NIPS'01, MIT Press, Cambridge, MA, USA (2001)
25. Perozzi, B., Al-Rfou, R., Skiena, S.: Deepwalk: Online learning of social representations. In: KDD. p. 701–710 (2014)

26. Qiu, J., Dong, Y., Ma, H., Li, J., Wang, C., Wang, K., Tang, J.: NetSMF: Large-scale network embedding as sparse matrix factorization. In: WWW (2019)
27. Qiu, J., Dong, Y., Ma, H., Li, J., Wang, K., Tang, J.: Network embedding as matrix factorization: Unifying DeepWalk, LINE, PTE, and Node2Vec. In: WSDM. pp. 459–467 (2018)
28. Raftery, A.E., Niu, X., Hoff, P.D., Yeung, K.Y.: Fast inference for the latent space network model using a case-control approximate likelihood. Journal of Computational and Graphical Statistics 21(4), 901–919 (2012)
29. Ryan, C., Wyse, J., Friel, N.: Bayesian model selection for the latent position cluster model for social networks. Network Science 5(1), 70–91 (2017)
30. Sun, B.J., Shen, H., Gao, J., Ouyang, W., Cheng, X.: A non-negative symmetric encoder-decoder approach for community detection. In: CIKM (2017)
31. Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., Mei, Q.: LINE: Large-scale information network embedding. In: WWW. pp. 1067–1077 (2015)
32. Wang, X., Cui, P., Wang, J., Pei, J., Zhu, W., Yang, S.: Community preserving network embedding. In: AAAI (2017)
33. Wind, D.K., Mørup, M.: Link prediction in weighted networks. In: 2012 IEEE Int. Workshop MLSP. pp. 1–6 (2012)
34. Yang, J., Leskovec, J.: Overlapping community detection at scale: A nonnegative matrix factorization approach. In: WSDM (2013)
35. Yang, L., Gu, J., Wang, C., Cao, X., Zhai, L., Jin, D., Guo, Y.: Toward unsupervised graph neural network: Interactive clustering and embedding via optimal transport. In: ICDM (2020)
36. Zhang, D., Yin, J., Zhu, X., Zhang, C.: Network representation learning: A survey. IEEE Trans. Big Data 6(1) (2020)
37. Zhang, J., Dong, Y., Wang, Y., Tang, J., Ding, M.: Prone: Fast and scalable network representation learning. In: IJCAI (2019)

Characterizing Polarization in Social Networks using the Signed Relational Latent Distance Model

Nikolaos Nakis
Technical University of Denmark

Abdulkadir Çelikkanat
Technical University of Denmark

Louis Boucherie
Technical University of Denmark

Christian Djurhuus
Technical University of Denmark

Felix Burmester
Technical University of Denmark

Daniel Mathias Holmelund
Technical University of Denmark

Monika Frolcová
Technical University of Denmark

Morten Mørup
Technical University of Denmark

Abstract

Graph representation learning has become a prominent tool for the characterization and understanding of the structure of networks in general and social networks in particular. Typically, these representation learning approaches embed the networks into a low-dimensional space in which the role of each individual can be characterized in terms of their latent position. A major current concern in social networks is the emergence of polarization and filter bubbles promoting a mindset of "us-versus-them" that may be defined by extreme positions believed to ultimately lead to political violence and the erosion of democracy. Such polarized networks are typically characterized in terms of signed links reflecting likes and dislikes. We propose the Signed Latent Distance Model (SLDM) utilizing for the first time the Skellam distribution as a likelihood function for signed networks. We further extend the modeling to the characterization of distinct extreme positions by constraining the embedding space to polytopes, forming the Signed Latent relational dIstance Model (SLIM). On four real social signed networks of polarization, we demonstrate that the models extract low-dimensional characterizations that well predict friendships and animosity while SLIM provides interpretable visualizations defined by extreme positions when restricting the embedding space to polytopes.

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

1 INTRODUCTION

For several decades, the origin and influence of political polarization have been issues receiving considerable attention both within scholarly research and the public media (Hetherington, 2009). Several studies have demonstrated an increasing partisan polarization among the political elites, some of which rely on network science approaches, for instance, using co-voting similarity networks and modularity to model and explain the distinct aspects of the data (Moody and Mucha, 2013). Whereas polarization has been described in terms of communities and their boundary properties (Guerra et al., 2013), latent distance modeling has also been used to extract bipolar structures (Barberá et al., 2015).

Ideological polarization is the distance between policy preferences, typically of elites taking extreme stands on issues whereas the electoral behavior is denoted affective polarization. When these extremes are portrayed as existential in the media, they typically form an "us-versus-them"-mindset (Dagnes, 2019). From a social network perspective, the process of polarization has been described to occur when "homophily and influence become self-reinforcing when the attraction to those who are similar and differentiation from those who are dissimilar entail greater openness to influence. The result is network autocorrelation—the tendency for people to resemble their network neighbors" (DellaPosta et al., 2015).

To better capture ideological polarization, we turn to signed networks. Signed networks reflect complex social polarization better than unsigned networks because they capture positive, negative, and neutral relationships between entities. The study of signed networks goes back to the '50s and was motivated by friendly and hostile social relationships (Harary, 1953). Since then they have been used to study networks of Twitter users (Keuchenius et al., 2021)

and US Congress members (Thomas et al., 2006), two examples of polarized social networks (Garimella and Weber, 2017; Neal, 2020).

In this paper, we focus on polarization as extreme positions and argue that the multi-polarity of "us-versus-them" reinforced by homophily and influence can be characterized by a latent position model (i.e., the latent distance model (Hoff et al., 2002)) of networks confined to a constrained social space formed by a polytope, what we denote a sociotope. As such, the corners of the sociotope define distinct aspects (i.e., poles) formed by polarized networks' tendencies to self-reinforce homophily by positive ties driving those who are similar close as opposed to those that are negatively tied being repelled. This can be revealed in terms of the important multiple poles of social network defining corners of such sociotope. Within these corners, positive interactions between nodes place them in close proximity in space thereby accounting for homophily while negative interactions "push" nodes far apart (towards opposing poles) yielding the "us-versus-them" effect.

The conceptual idea of polytopes as formed by pure types can be traced back to Plato's forms, which characterize the physical world as a limited projection of the forms also referred to as ideal categories. Later, Carl Jung introduced the concept of universal archetypes, described as a collective unconscious, in which he related to Plato's forms by describing the forms as a Jungian version of the Platonian archetypes (Williamson, 1985). Employing the theoretical concept of archetypes to political and ideological polarization, the archetypes could be interpreted as genuine ideologies, while the ideological advocates can be expressed as a mixture of distinct ideologies.

Archetypal Analysis (AA) is a prominent framework for extracting polytopes in tabular data. AA was originally proposed by Cutler and Breiman (1994) as an unsupervised learning method that favors distinct aspects, archetypes, of the data in which observations are characterized by convex combinations (i.e., mixtures) of these archetypes as opposed to clustering procedures extracting prototypical observations (Mørup and Kai Hansen, 2010). AA has previously been used to model societal conflicts in Europe (Beugelsdijk et al., 2022). However, given that AA was proposed for tabular data, the applications are currently restricted to non-relational data. Thus, whereas the characterization of data in terms of distinct aspects and polytopes has a long history, such representation learning approaches have not previously been considered in the context of network analysis for the extraction of polarization by several extremes.

In the last years, representation learning of signed graphs has gathered substantial attention, with applications such as signed link prediction (Chiang et al., 2011), and community detection (Tzeng et al., 2020). Initial works extended the prominent random walks framework (Perozzi et al., 2014; Grover and Leskovec, 2016) to the analysis of signed networks. SIDE (Kim et al., 2018b) exploits truncated random walks on the signed graph with interaction signs for each node pair inferred based on balance theory (Cartwright and Harary, 1956). Balance theory is a socio-psychological theory admitting four rules: "The friend of my friend is my friend," "The friend of my enemy is my enemy," "The enemy of my friend is my enemy," and "The enemy of my enemy is my friend." POLE (Huang et al., 2022), also utilizes balance theory-based signed random walks to construct an auto-covariance similarity which is used to obtain the embedding space. Neural networks have also been adopted for the analysis of signed networks. Both SINE (Wang et al., 2017) and SIGNET (Islam et al., 2018) combine balance theory and multi-layer neural networks to learn the network embeddings while SIGNet uses targeted node sampling to provide scalable inference. In addition, graph neural networks have also been studied in the context of signed graphs. More specifically, SIGAT (Huang et al., 2019) and SDGNN (Huang et al., 2021) combine balance and status theory with graph attention to learn signed network embeddings. The status theory is another important socio-psychological theory for directed relationships where for a source and a target node, a positive directed connection assumes a higher status of the target, i.e. $\{\text{status(target)} > \text{status(source)}\}$, while the inequality is opposite for a negative connection. Lastly, SLF (Xu et al., 2019) learns multiple latent factors of the signed network, modeling positive, negative, and neutral, as well as the absence of a relationship between node pairs.

A prominent approach for graph representation learning is the Latent Distance Model (Hoff et al., 2002) in which the tendency of nodes to connect is defined in terms of their proximity in latent space. Notably, the LDM can express the properties transitivity (*"a friend of a friend is a friend"*) and homophily (*"akin nodes tend to have links"*). Recently, it has been shown that LDMs can account for the structure of networks in ultra-low dimensions (Nakis et al., 2022, 2023; Çelikkanat et al., 2022). It has further been demonstrated that an LDM of one dimension can be used to extract bipolar network properties (Barberá et al., 2015).

For the modeling of signed networks for the characterization of polarization, we first present the Signed Latent Distance Model (SLDM). The model utilizes a likelihood function for weighted signed links based on the Skellam distribution (Skellam, 1946). The Skellam distribution is the discrete probability distribution of the difference between two independent Poisson random variables. It was introduced by John Gordon Skellam to model the dynamics of populations (Skellam, 1946). Since then it was used in medicine to model treatment measurements (Karlis and Ntzoufras, 2006), sports results (Karlis and Ntzoufras, 2008), as well as, econometric studies (Barndorff-Nielsen et al.,

2010). Furthermore, we introduce the Signed relational Latent dIstance Model (SLIM) being able to characterize the latent social space in terms of extreme positions forming polytopes inspired by archetypal analysis enabling archetypal analysis for relational data, i.e. relational AA (RAA). We apply SLDM and SLIM on four real signed networks believed to reflect polarization and demonstrate how SLIM uncovers prominent distinct positions (poles). To the best of our knowledge, this is the first work to model signed weighted networks using the Skellam distribution and the first time AA has been extended to relational data by leveraging latent position modeling approaches for the characterization of polytopes in social networks. **The implementation is available at:** *github.com/Nicknakis/SLIM_RAA*.

2 PROPOSED METHODOLOGY

Let $\mathcal{G} = (\mathcal{V}, \mathcal{Y})$ be a *signed graph* where $\mathcal{V} = \{1, \dots, N\}$ denotes the set of vertices and $\mathcal{Y} : \mathcal{V}^2 \to \mathsf{X} \subseteq \mathbb{R}$ is a map indicating the weight of node pairs, such that there is an edge $(i, j) \in \mathcal{V}^2$ if the weight $\mathcal{Y}(i, j)$ is different from 0. In other words, $\mathcal{E} := \{(i, j) \in \mathcal{V}^2 : \mathcal{Y}(i, j) \neq 0\}$ indicates the set of edges of the network. Since many real networks consist of only integer-valued edges, in this paper, we set X to $\mathbb{Z}$, and we will call the graph *undirected* if the pairs (i, j) and (j, i) represent the same link. (The directed case is provided in the supplementary materials.) For simplicity, y_{ij} denotes each edge weight.

2.1 The Skellam Latent Distance Model (SLDM)

Our main purpose is to learn latent node representations $\{\mathbf{z}_i\}_{i\in\mathcal{V}} \in \mathbb{R}^K$ in a low dimensional space for a given signed network $\mathcal{G} = (\mathcal{V}, \mathcal{Y})$ ($K \ll |\mathcal{V}|$). Therefore, the edge weights can take any integer value to represent the positive or negative tendencies between the corresponding nodes. We model these signed interactions among the nodes using the Skellam distribution (Skellam, 1946), which can be formulated as the difference of two independent Poisson-distributed random variables ($y = N_1 - N_2 \in \mathbb{Z}$) with respect to the rates λ^+ and λ^-:

$$P(y|\lambda^+, \lambda^-) = e^{-(\lambda^+ + \lambda^-)} \left(\frac{\lambda^+}{\lambda^-}\right)^{y/2} \mathcal{I}_{|y|}\left(2\sqrt{\lambda^+\lambda^-}\right),$$

where $N_1 \sim Pois(\lambda^+)$ and $N_2 \sim Pois(\lambda^-)$, and $\mathcal{I}_{|y|}$ is the modified Bessel function of the first kind and order $|y|$. To the best of our knowledge, the Skellam distribution has not been adapted before for modeling the network likelihood. More specifically, we propose a novel latent space model utilizing the Skellam distribution by adopting the latent distance model, which was proposed originally for undirected, and unsigned binary networks as a logistic regression model (Hoff et al., 2002). It was later extended to multiple generalized linear models (Hoff, 2005), including the Poisson regression model for integer-weighted networks. We can formulate the negative log-likelihood of a latent distance model under the Skellam distribution as:

$$\begin{aligned}\mathcal{L}(\mathcal{Y}) &:= \log p(y_{ij}|\lambda_{ij}^+, \lambda_{ij}^-) \\ &= \sum_{i<j} (\lambda_{ij}^+ + \lambda_{ij}^-) - \frac{y_{ij}}{2} \log\left(\frac{\lambda_{ij}^+}{\lambda_{ij}^-}\right) - \log(I_{ij}^*),\end{aligned}$$

where $I_{ij}^* := \mathcal{I}_{|y_{ij}|}\left(2\sqrt{\lambda_{ij}^+\lambda_{ij}^-}\right)$. As it can be noticed, the Skellam distribution has two rate parameters, and we consider them to learn latent node representations $\{\mathbf{z}_i\}_{i\in\mathcal{V}}$ by defining them as follows:

$$\lambda_{ij}^+ = \exp\left(\gamma_i + \gamma_j - ||\mathbf{z}_i - \mathbf{z}_j||_2\right), \tag{1}$$

$$\lambda_{ij}^- = \exp\left(\delta_i + \delta_j + ||\mathbf{z}_i - \mathbf{z}_j||_2\right), \tag{2}$$

where the set $\{\gamma_i, \delta_i\}_{i\in\mathcal{V}}$ denote the node-specific random effect terms, and $||\cdot||_2$ is the Euclidean distance function. More specifically, γ_i, γ_j represent the "social" effects/reach of a node and the tendency to form (as a receiver and as a sender, respectively) positive interactions, expressing positive degree heterogeneity (indicated by $+$ as a superscript of λ). In contrast, δ_i, δ_j provide the "anti-social" effect/reach of a node to form negative connections, and thus models negative degree heterogeneity (indicated by $-$ as a superscript of λ).

By imposing standard normally distributed priors elementwise on all model parameters $\boldsymbol{\theta} = \{\boldsymbol{\gamma}, \boldsymbol{\delta}, \mathbf{Z}\}$, i.e., $\theta_i \sim \mathcal{N}(0, 1)$, We define a maximum a posteriori (MAP) estimation over the model parameters, via the loss function to be minimized (ignoring constant terms):

$$\begin{aligned}Loss = &\sum_{i<j} \left(\lambda_{ij}^+ + \lambda_{ij}^- - \frac{y_{ij}}{2} \log\left(\frac{\lambda_{ij}^+}{\lambda_{ij}^-}\right)\right) \\ &- \sum_{i<j} \log I_{|y_{ij}|}\left(2\sqrt{\lambda_{ij}^+\lambda_{ij}^-}\right) \\ &+ \frac{\rho}{2}\Big(||\mathbf{Z}||_F^2 + ||\boldsymbol{\gamma}||_F^2 + ||\boldsymbol{\delta}||_F^2\Big),\end{aligned} \tag{3}$$

where $||\cdot||_F$ denotes the Frobenius norm. In addition, ρ is the regularization strength with $\rho = 1$ yielding the adopted normal prior with zero mean and unit variance. Importantly, by setting λ_{ij}^+ and λ_{ij}^- based on Eq. (11) and (2), the model effectively makes positive (weighted) links attract and negative (weighted links) deter nodes from being in proximity of each other.

2.2 Archetypal Analysis

Archetypal Analysis (AA) (Cutler and Breiman, 1994; Mørup and Kai Hansen, 2010) is an approach developed for the modeling of observational data in which the data is expressed in terms of convex combinations of characteristics (i.e. archetypes). The definition of the embedded data

points is given as follows:

$$\mathbf{X} \approx \mathbf{XCZ} \quad \text{s.t. } \boldsymbol{c}_d \in \Delta^N \text{ and } \mathbf{z}_j \in \Delta^K \tag{4}$$

where Δ^P denotes the standard simplex in $(P+1)$ dimensions such that $\mathbf{q} \in \Delta^P$ requires $q_i \geq 0$ and $\|\mathbf{q}\|_1 = 1$ (i.e. $\sum_i q_i = 1$). Notably, the archetypes given by the columns of $\mathbf{A} = \mathbf{XC}$ define the corners of the extracted polytope as convex combinations of the observations, whereas $\mathbf{Z}$ define how each observation is reconstructed as convex combinations of the extracted archetypes.

Whereas archetypal analysis constrains the representation to the convex hull of the data, other approaches to model pure/ideal forms have been Minimal Volume (MV) approaches defined by

$$\mathbf{X} \approx \mathbf{AZ} \quad \text{s.t. } vol(\mathbf{A}) = v \text{ and } \mathbf{z}_j \in \Delta^K, \tag{5}$$

in which $vol(\mathbf{A})$ defines the volume of $\mathbf{A}$. When $\boldsymbol{A}$ is a square matrix this can be defined by $vol(\mathbf{A}) = |det(\mathbf{A})|$, see also Hart et al. (2015); Zhuang et al. (2019) for a review on such end-member extraction procedures. A strength is that, as opposed to AA, the approach does not require the presence of pure observations, however, a drawback is a need for regularization tuning to define an adequate volume (Zhuang et al., 2019) whereas the exact computation of the volume of general polytopes requires the computation of determinants of the sum of all simplices defining the polytope (Büeler et al., 2000). Importantly, Archetypal Analysis and Minimal volume extraction procedures have been found to identify latent polytopes defining trade-offs in which vertices of the polytopes represent maximally enriched distinct aspects (archetypes), allowing identification of tasks or prominent roles the vertices of the polytope represent (Shoval et al., 2012; Hart et al., 2015). Due to the computational issues of regularizing high-dimensional volumes and the need for careful tuning of such regularization parameters, we presently focus on polytope extraction as defined through the AA formulation rather than the MV formulation.

2.3 A Generative Model of Polarization

Considering a latent space for the modeling of polarization, we presently extend the Skellam LDM and define polarization as extreme positions (pure forms/archetypes) that optimally represent the social dynamics observed in terms of the induced polytope - what we denote a sociotope, in which each observation is a convex combination of these extremes. In particular, we characterize polarization in terms of extreme positions in a latent space defined as a polytope akin to AA and MV.

In our generative model of polarization, we further suppose that the bias terms introduced in the definitions of the Poisson rates, $(\lambda_{ij}^+, \lambda_{ij}^-)$, are normally distributed. Since latent representations $\{\mathbf{z}_i\}_{i \in \mathcal{V}}$ according to AA and MV lie in the standard simplex set Δ^K, we further assume that they follow a Dirichlet distribution. Formally, we can summarize the generative model as follows:

$$\begin{aligned}
\gamma_i &\sim \mathcal{N}(\mu_\gamma, \sigma_\gamma^2) && \forall i \in \mathcal{V}, \\
\delta_i &\sim \mathcal{N}(\mu_\delta, \sigma_\delta^2) && \forall i \in \mathcal{V}, \\
\mathbf{a}_k &\sim \mathcal{N}(\boldsymbol{\mu}_A, \sigma_A^2 \mathbf{I}) && \forall k \in \{1, \ldots, K\}, \\
\mathbf{z}_i &\sim Dir(\boldsymbol{\alpha}) && \forall i \in \mathcal{V}, \\
\lambda_{ij}^+ &= \exp\left(\gamma_i + \gamma_j - \|\mathbf{A}(\mathbf{z}_i - \mathbf{z}_j)\|_2\right), \\
\lambda_{ij}^- &= \exp\left(\delta_i + \delta_j + \|\mathbf{A}(\mathbf{z}_i - \mathbf{z}_j)\|_2\right), \\
y_{ij} &\sim Skellam(\lambda_{ij}^+, \lambda_{ij}^-) && \forall (i,j) \in \mathcal{V}^2.
\end{aligned}$$

According to the above generative process, positive ($\boldsymbol{\gamma}$) and negative ($\boldsymbol{\delta}$) random effects for the nodes are first drawn, upon which the location of extreme positions $\mathbf{A}$ (i.e., corners of the polytope denoted archetypes) are generated. In addition, as the dimensionality of the latent space increases linearly with the number of archetypes, i.e. $\boldsymbol{A}$ is a square matrix, with probability zero archetypes will be placed in the interior of the convex hull of the other archetypes. Subsequently, the node-specific convex combinations $\mathbf{Z}$ of the generated archetypes are drawn, and finally, the weighted signed link is generated according to the node-specific biases and distances between dyads within the polytope utilizing the Skellam distribution.

2.4 The Signed Relational Latent Distance Model

For inference, we exploit how polytopes can be efficiently extracted using archetypal analysis. We, therefore, define the Signed Latent relational dIstance Model (SLIM) by defining a relational archetypal analysis approach endowing the generative model a parameterization akin to archetypal analysis in order to efficiently extract polytopes from relational data defined by signed weighted networks. Specifically, we formulate the relational AA in the context of the family of LDMs, as:

$$\begin{aligned}
\lambda_{ij}^+ &= \exp\left(\gamma_i + \gamma_j - \|\mathbf{A}(\mathbf{z}_i - \mathbf{z}_j)\|_2\right) && (6) \\
&= \exp\left(\gamma_i + \gamma_j - \|\mathbf{RZC}(\mathbf{z}_i - \mathbf{z}_j)\|_2\right). && (7) \\
\lambda_{ij}^- &= \exp\left(\delta_i + \delta_j + \|\mathbf{A}(\mathbf{z}_i - \mathbf{z}_j)\|_2\right) && (8) \\
&= \exp\left(\delta_i + \delta_j + \|\mathbf{RZC}(\mathbf{z}_i - \mathbf{z}_j)\|_2\right). && (9)
\end{aligned}$$

Notably, in the AA formulation $\mathbf{X} = \mathbf{RZ}$ corresponds to observations formed by convex combinations $\mathbf{Z}$ of positions given by the columns of $\boldsymbol{R}^{K \times K}$. Furthermore, in order to ensure what is used to define archetypes $\mathbf{A} = \mathbf{XC} = \mathbf{RZC}$ corresponds to observations using these archetypes in their reconstruction $\mathbf{Z}$, we define $\boldsymbol{C} \in \boldsymbol{R}^{N \times K}$ as a gated version of $\mathbf{Z}$ normalized to the simplex such that $\boldsymbol{c}_d \in \Delta^N$ by defining

$$c_{nd} = \frac{(\mathbf{Z}^\top \circ [\sigma(\mathbf{G})]^\top)_{nd}}{\sum_{n'} (\mathbf{Z}^\top \circ [\sigma(\mathbf{G})]^\top)_{n'd}} \tag{10}$$

in which $\circ$ denotes the elementwise (Hadamard) product and $\sigma(\mathbf{G})$ defines the logistic sigmoid elementwise applied to the matrix $\boldsymbol{G}$. As a result, the extracted archetypes are ensured to correspond to the nodes assigned the archetype, whereas the location of the archetypes can be flexibly placed in space as defined by $\mathbf{R}$. By defining $\mathbf{z}_i = \mathrm{softmax}(\tilde{\mathbf{z}}_i)$ we further ensure $\mathbf{z}_i \in \Delta^K$.

Importantly, the loss function of Eq. (13) is adopted for the relational AA formulation forming the SLIM, with the prior regularization applied to the corners of the extracted polytope $\boldsymbol{A} = \mathbf{RZC}$ instead of the latent embeddings $\boldsymbol{Z}$ imposing a standard elementwise normal distribution as prior $a_{k,k'} \sim \mathcal{N}(0,1)$. Furthermore, we impose a uniform Dirichlet prior on the columns of $\boldsymbol{Z}$, i.e. $(\boldsymbol{z}_i \sim Dir(\mathbf{1}_K)$, this only contributes constant terms to the joint distribution, and therefore the maximum a posteriori (MAP) optimization only constant terms. As a result, the loss function optimized is given by Eq. (13) replacing $\|\mathbf{Z}\|_F^2$ with $\|\boldsymbol{A}\|_F^2$.

Complexity analysis. With SLDM/SLIM being distance models, they scale prohibitively as $\mathcal{O}(N^2)$ since the node pairwise distance matrix needs to be computed. This does not allow the analysis of large-scale networks. For that, we adopt an unbiased estimation of the log-likelihood through random sampling. More specifically, gradient steps are based on the log-likelihood of the block formed by a sampled (per iteration and with replacement) set S of network nodes. This makes inference scalable defining an $\mathcal{O}(S^2)$ space and time complexity. More options for scalable inference of distance models have also been proposed in Nakis et al. (2022); Raftery et al. (2012).

3 RESULTS AND DISCUSSION

We extensively evaluate the performance of our proposed methods by comparing them to the prominent GRL approaches designed for signed networks. All experiments regarding SLDM/SLIM have been conducted on an 8 GB NVIDIA RTX 2070 Super GPU. In addition, we adopted the Adam optimizer Kingma and Ba (2017) with learning rate $\mathrm{lr} = 0.05$ and for 5000 iterations. The sample size for the node set was chosen as approximately 3000 nodes for all networks. The initialization of the SLDM/SLIM frameworks is deterministic and based on the spectral decomposition of the normalized Laplacian (more details are provided in the supplementary).

Artificial networks. We first, introduce experiments on artificial networks, as generated by the generative process described in Section 2.3. We create two networks expressing different levels of polarization. Results are presented in Fig. 1. More specifically, sub-Figs 1a and 1e show the ground truth latent spaces generating the networks with adjacency matrices as shown by sub-Figs 1b and 1f, respectively. The inferred latent spaces of the two networks are provided in sub-Figs 1c and 1g where it is clear that the model successfully distinguishes the difference in the level of polarization of the two networks. We also verify the generated networks based on the inferred parameters given by sub-Figs 1d and 1h. We observe that the model successfully generates sparse networks accounting for the positive and negative link imbalance.

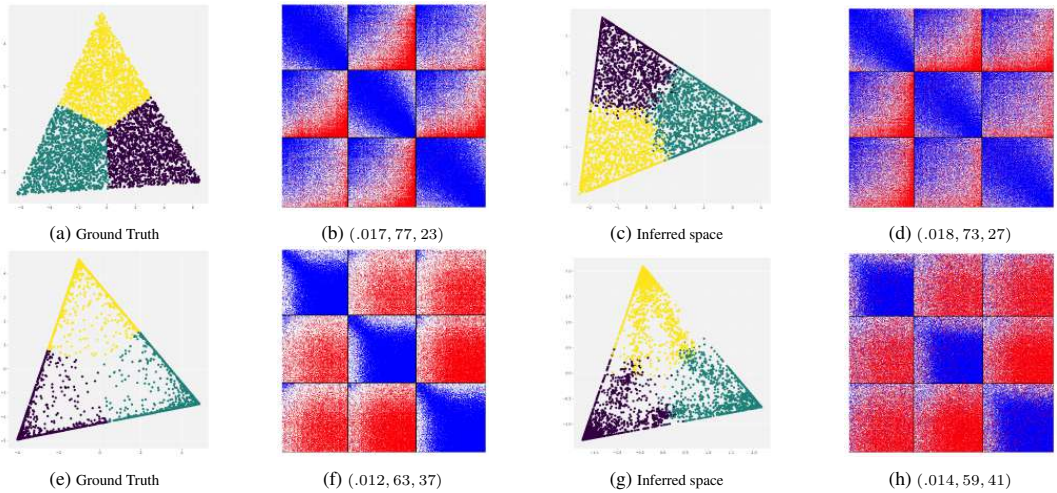
Table 1: Network statistics; $|\mathcal{V}|$: # Nodes, $|\mathcal{Y}^+|$: # Positive links, $|\mathcal{Y}^-|$: # Negative links.

	$\lvert\mathcal{V}\rvert$	$\lvert\mathcal{Y}^+\rvert$	$\lvert\mathcal{Y}^-\rvert$	Density
Reddit	35,776	128,182	9,639	0.0001
Twitter	10,885	238,612	12,794	0.0021
wiki-Elec	7,117	81,277	21,909	0.0020
wiki-RfA	11,332	117,982	66,839	0.0014

Real networks. We employed four networks of varying sizes and structures. (**i**) *Reddit* is constructed based on hyperlinks representing the directed connections between two communities in a social platform (Kumar et al., 2018). (**ii**) *wikiRfA* and (**iii**) *wikiElec* are the election networks covering the different time intervals in which nodes indicate the users and the directed links show supporting, neutral, and opposing votes to be selected as an administrator on the Wikipedia platform (West et al., 2014; Leskovec et al., 2010). Finally, (**iv**) *Twitter* is an undirected social network built on the corpus of tweets concerning the highly polarized debate about the reform of the Italian Constitution (Ordozgoiti et al., 2020).

In our experiments, we consider the greatest connected component of the networks, and if the original network is temporal, we construct the static network by summing the weights of the links through time. For the experiments performed on undirected graphs, we similarly combine directed links to obtain the undirected version of the networks.

Baselines. We benchmark the performance of our proposed frameworks against five prominent graph representation learning methods, designed for the analysis of signed networks: (**i**) POLE (Huang et al., 2022) which learns the network embeddings by decomposing the signed random walks auto-covariance similarity matrix, (**ii**) SLF (Xu et al., 2019) learns embeddings that are the concatenation of two latent factors targeting positive and negative relations, (**iii**) SIGAT (Huang et al., 2019) is a graph neural network approach using graph attention to learn node embeddings, (**iv**) SIDE (Kim et al., 2018b) is another random walk based method for signed networks, (**v**) SIGNET (Islam et al., 2018) is a multi-layer neural network approach constructing a Hadamard product similarity to accommodate for signed proximity on the network pairwise relations.



(a) Ground Truth (b) (.017, 77, 23) (c) Inferred space (d) (.018, 73, 27)

(e) Ground Truth (f) (.012, 63, 37) (g) Inferred space (h) (.014, 59, 41)

Figure 1: Two artificially generated networks with different levels of polarization $\{\mathbf{z}_i \sim Dir(\mathbf{1})$ (top row), and $\mathbf{z}_i \sim Dir(0.1 \cdot \mathbf{1})$ (bottom row)$\}$. Both size $N = 5000$ nodes and $K = 3$ archetypes. The first column shows the first two principal components of the original latent space $\tilde{\boldsymbol{Z}} = \boldsymbol{AZ}$, the second column the original adjacency matrix, while the parenthesis shows the network statistics as: (density,% of positive (blue) links,% of negative (red) links). The third column displays the first two principal components of the inferred latent space, and the fourth column is the SLIM generated network based on inferred parameters. All network adjacency matrices are ordered based on $\mathbf{z}_i$, in terms of maximum archetype membership and internally according to the magnitude of the corresponding archetype most used for their reconstruction.

Table 2: Area Under Curve (AUC-ROC) scores for representation size of $K = 8$.

	WikiElec			*WikiRfa*			*Twitter*			*Reddit*		
Task	$p@n$	$p@z$	$n@z$	$p@n$	$p@z$	$n@z$	$p@n$	$p@z$	$n@z$	$p@n$	$p@z$	$n@z$
POLE	.809	.896	.853	.904	.921	.767	.965	.902	.922	x	x	x
SLF	**.888**	.954	**.952**	**.971**	.963	.961	.914	.877	.968	**.729**	**.955**	.968
SIGAT	.874	.775	.754	.944	.766	.792	**.998**	.875	.963	.707	.682	.712
SIDE	.728	.866	.895	.869	.861	.908	.799	.843	.910	.653	.830	.892
SIGNET	.841	.774	.635	.920	.736	.717	.968	.719	.891	.646	.547	.623
SLIM (OURS)	.862	.965	.935	.956	.980	.960	.988	**.963**	**.972**	.667	**.955**	**.978**
SLDM (OURS)	.876	**.969**	.936	.963	**.982**	**.963**	.986	.962	**.973**	.648	.951	.975

Table 3: Area Under Curve (AUC-PR) scores for representation size of $K = 8$.

	WikiElec			*WikiRfa*			*Twitter*			*Reddit*		
Task	$p@n$	$p@z$	$n@z$	$p@n$	$p@z$	$n@z$	$p@n$	$p@z$	$n@z$	$p@n$	$p@z$	$n@z$
POLE	.929	.922	.544	.927	.937	.779	.998	.932	.668	x	x	x
SLF	**.964**	.926	**.787**	**.983**	.922	.881	.994	.870	.740	**.966**	.956	**.850**
SIGAT	.960	.724	.439	.969	.646	.497	**.999**	.861	.582	.965	.692	.232
SIDE	.907	.779	.608	.920	.806	.739	.974	.831	.469	.957	.820	.614
SIGNET	.944	.670	.298	.950	.572	.417	.998	.647	.248	.956	.510	.083
SLIM (OURS)	.953	.956	.785	.973	.969	.907	**.999**	.962	**.813**	.958	**.960**	**.850**
SLDM (OURS)	.960	**.963**	**.787**	.977	**.971**	**.912**	**.999**	**.963**	.809	.954	.955	.846

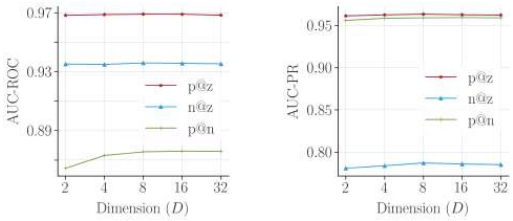
3.1 Link prediction

We evaluate performance considering the link prediction task considering the ability of our model to predict links of disconnected network pairs which should be connected, as well as, infer the sign of these links (positive or negative). For this, we remove/hide 20% of the total network links while preserving connectivity on the residual network. For the testing set, the removed edges are paired with a sample of the same number of node pairs that are not the edges of the original network to create zero instances. To learn the node embeddings, we make use of the residual network.

Predictions and evaluation metrics. For our methods we fit a logistic regression classifier on the concatenation of the corresponding Skellam rates and log-rates, as $\chi_{ij} = [\lambda^+_{ij}, \lambda^-_{ij}, \log \lambda^+_{ij}, \log \lambda^-_{ij}]$. Since our Skellam likelihood formulation relies both on the ratio and products of the rates, a concatenation can take advantage of a linear function of the rates, as well as, their ratio or product as allowed from the log transformation. For the baselines, we use five binary operators {average, weighted L1, weighted L2, concatenate, Hadamard product} to construct feature vectors. For each of these feature vectors, we fit a logistic regression model (except for the Hadamard product which is used directly for predictions). Since different operators provide different performances, for the baselines we choose the operator that returns the maximum performance per individual task. As a consequence of the class imbalances and the sparsity present in signed networks, we adopt robust evaluation metrics, such as area-under-curve of the receiver operating characteristic (AUC-ROC) and precision-recall (AUC-PR) curves. Lastly, we denote with "x" the performance of a baseline if it was unable to run due to high memory/runtime complexity.

Link sign prediction. In this setting, we utilize the link test set containing the negative/positive cases of removed connections. We then ask the models to predict the sign of the removed links. We denote the task of the link sign prediction task as $p@n$. In Table 2 we provide the AUC-ROC scores while in Table 3 the AUC-PR scores for the undirected case. Here we observe that our proposed models outperform the baselines in most networks while being competitive in the *Reddit* network against SLF. This specific baseline is the most competitive across networks showing high and consistent performance similar to SLIM and SLDM. Comparing now SLIM with SLDM we get mostly on-par results, verifying that constraining the model to a polytope still provides enough expressive capability as the unconstrained model while allowing for accurate extraction of "extreme" positions.

Signed link prediction. The second and more challenging task is to predict removed links against disconnected pairs of the network, as well as, infer the sign of each link correctly. For that, the test set is split into two subsets positive/disconnected and negative/disconnected. We then evaluate the performance of each model on those subsets. The tasks of signed link prediction between positive and zero samples are denoted as $p@z$ while the negative against zero is $n@z$. We summarize our results by presenting AUC-ROC and AUC-PR scores in Table 2 and Table 3 respectively. Once more our models outperform the baselines in most networks and for both versions of signed link prediction. The SLF baseline is again the most competitive baseline yielding on-par results for *Reddit*.



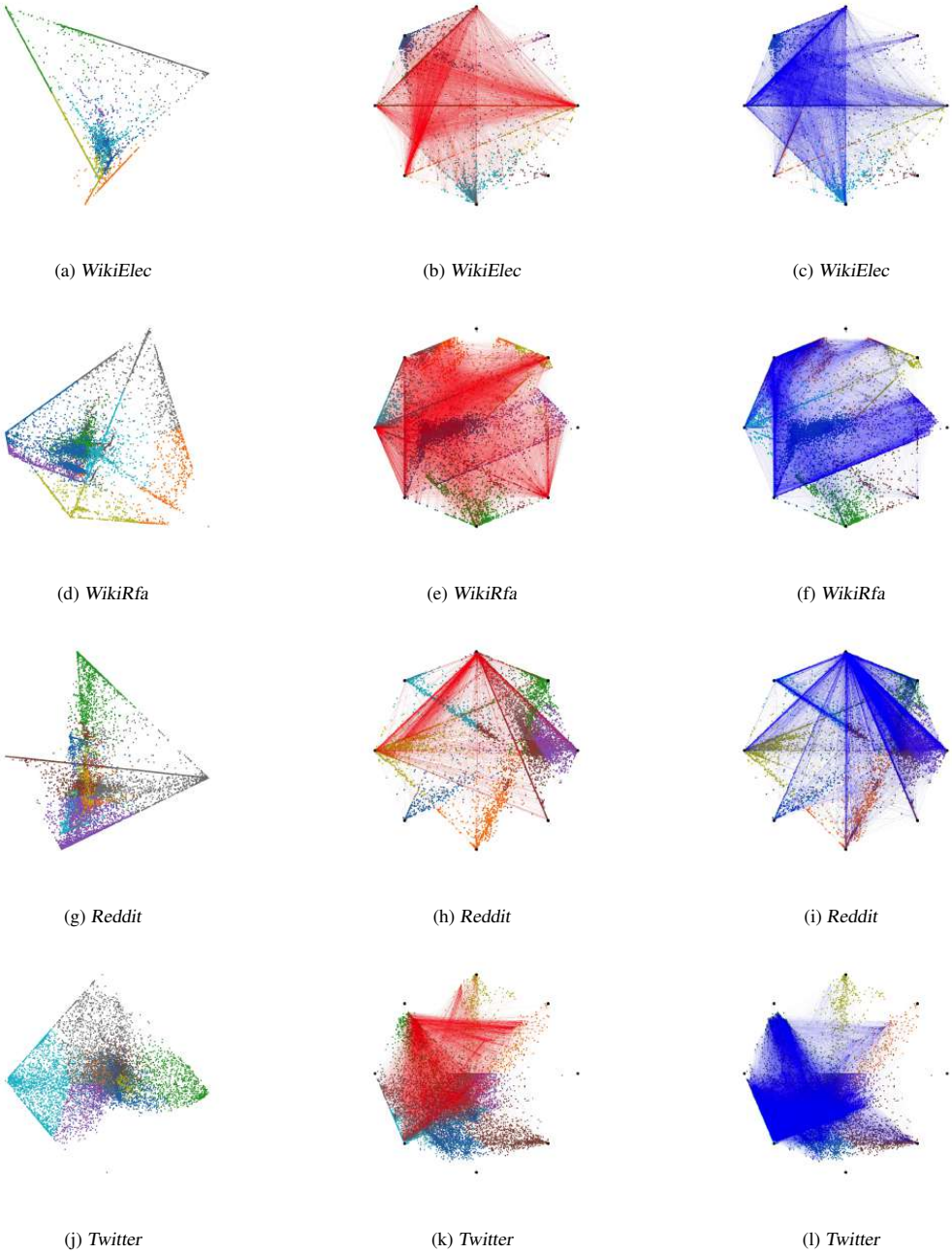
(a) (b)

Figure 2: *wikiElec*: Performance of SLIM across dimensions for different tasks, (a) Area-Under-Curve Receiver Operating Characteristic scores, (b) Area-Under-Curve Precision-Recall scores. Both AUC-ROC and AUC-PR scores are almost constant across different dimensions

Directed networks. Directed network results are provided in the supplementary. Since SLF has higher modeling capacity it outperforms the simple model formulation of SLDM and SLIM. For that, we explore and discuss formulations allowing for more capacity in the SLDM/SLIM model for the directed case (see supplementary).

Effect of dimensionality. In Figure 2, we provide the performance across dimensions for the different downstream task and for the *wikiElec* dataset. We observe that both AUC-ROC and AUC-PR scores are almost constant across different dimensions (note that as $R^{K\times K}$ dimensions for the SLIM is given by the number of archetypes), showcasing that increasing the models' capacity (in terms of dimensions) does not have a significant effect on the performance of these downstream tasks (similar results were observed for all networks and most of the baselines).

Visualizations. The RAA formulation facilitates the inference of a polytope describing the distinct aspects of networks. Here, we visualize the latent space across $K = 8$ dimensions for all of the corresponding networks. To facilitate visualizations we use Principal Component Analysis (PCA), and project the space based on the first two principal components of the final embedding matrix $\tilde{\boldsymbol{Z}} = \boldsymbol{AZ}$. In addition, we provide circular plots where each archetype of the polytope is mapped to a circle every $\text{rad}_k = \frac{2\pi}{K}$ radians, with K being the number of archetypes. Figure 3 contains three columns with the first denoting the PCA-



(a) *WikiElec* (b) *WikiElec* (c) *WikiElec*

(d) *WikiRfa* (e) *WikiRfa* (f) *WikiRfa*

(g) *Reddit* (h) *Reddit* (i) *Reddit*

(j) *Twitter* (k) *Twitter* (l) *Twitter*

Figure 3: Inferred polytope visualizations for various networks. The first column showcases the $K = 8$ dimensional sociotope projected on the first two principal components (PCA) — second and third columns provide circular plots of the sociotope enriched with the negative (red) and positive (blue) links, respectively.

induced space while the second and third columns correspond to the circular plots enriched by the negative (red) and positive (blue) links, respectively. We observe how the polytope successfully uncovers extreme positional nodes. More specifically, all networks have at least one archetype which acts as a "dislike" hub and at least one as a "like" hub. Meaning that these archetypes contain high values of negative/positive interactions. For the *wiki-RfA* and *Twitter* networks we observe archetypes of very low degree, this is explained due to some only "disliked" nodes being pushed away from the main node population. These can be regarded as "outliers" of the sociotope. Nevertheless, such outliers are discovered since they provide high expressive power for the model.

Discussion. The Signed Relational Latent Distance Model has been presented for the undirected case setting, and we employed the Euclidean distance for both Skellam rates λ_{ij}^{+}, λ_{ij}^{-}. The capacity of the current formulation works well for undirected networks. Nevertheless, there are alternative model formulations, and keeping the distance identical for the positive and negative rates constrains the models' expressive capability, especially for the directed/bipartite signed network case. We therefore explore additional model formulations such as setting the Skellam rates as, $\lambda_{ij}^{+} = \exp(\beta_i + \beta_j - ||\mathbf{z}_i - \mathbf{w}_j||_2)$ and $\lambda_{ij}^{-} = \exp(\gamma_i + \gamma_j - ||\boldsymbol{u}_i - \boldsymbol{w}_j||_2)$ in the supplementary material. Under this assumption, a positive directed relationship $(i \rightarrow j)$ shows that node i "likes" node j and "dislikes" node j if it is negative. The latent embedding $\boldsymbol{w}_j$ is then the receiver position for the "likes" and "dislikes" with embeddings $\mathbf{z}_i$ and $\boldsymbol{u}_i$ being the sender positions for positive and negative relationships, respectively. In this case, we introduce three latent embeddings instead of the conventional two for the undirected case. The disparity of location $\mathbf{z}_i$ and $\boldsymbol{u}_i$ here can point out how polarity is formed between the two regions of the latent space (Please see the supplementary material for further discussion and results).

Another important design characteristic for the SLDM/SLIM frameworks is the choice of the prior/regularization of the different parameters. So far, we did not tune any regularization strength of the priors and simply adopted a normal distribution on the model parameters and non-informative uniform Dirichlet prior on $\mathbf{Z}$ in the case of SLIM. Potential tuning of the priors with cross-validation is expected to boost the performance and results.

A prominent characteristic of signed networks is the sparsity or, in other words, the excess of "zero" weights among node pairs. An intriguing direction to account for it might be the zero-inflated version of the Skellam distribution (Karlis and Ntzoufras, 2008). Here essentially, we can define a mixture model responsible for the imbalance between cases (sign-weighted links) and controls (neutral zero links) in the network. Such zero-inflated SLDM/SLIM models can thereby define a generative process that can straightforwardly address different levels of network sparsity.

Whereas we consider the generalization of SLDM and SLIM to directed networks in the supplementary, a possible future direction should consider generalizations to bipartite networks in which we expect the directed generalizations to be applicable (Kim et al., 2018a; Nakis et al., 2022). Furthermore, networks of polarization typically evolve over time. Future work should thus investigate how the proposed modeling framework can be extended to characterize dynamic networks leveraging existing works by exploring dynamic extensions of latent space modeling approaches, including the diffusion model of (Sarkar and Moore, 2005) and approaches reviewed in Kim et al. (2018a).

4 CONCLUSION AND LIMITATIONS

The proposed Skellam Latent Distance Model (SLDM) and Signed Latent Relational Distance model (SLIM) provide easily interpretable network visualization with favorable performance in the link prediction tasks for weighted signed networks. In particular, endowing the model with a space constrained to polytopes (forming the SLIM) enabled us to characterize distinct aspects in terms of extreme positions in the social networks akin to conventional archetypal analysis but for graph-structured data. The Skellam distribution is considerably beneficial in modeling signed networks, whereas the relational extension of AA can be applied for other likelihood specifications, such as LDMs in general. This work thereby provides a foundation for using likelihoods accommodating weighted signed networks and representations akin to AA in general for analyzing networks.

The optimization for the SLDM/SLIM frameworks is a highly non-convex problem and thus relies on the quality of initialization in terms of convergence speed. In this regard, we use a deterministic initialization based on the normalized Laplacian. In addition, we observed that a maximum likelihood estimation of the model parameters became unstable when the network contained some nodes having only negative interactions. This is a direct consequence of the presence of the distance term $(\exp(+||\cdot||_2))$ for negative interactions, which can lead to overflow during inference. Nevertheless, the adopted MAP estimation was found to be stable across all networks. For real networks, the generative model created an "excess" of negative links increasing the overall network sparsity. For that, a modified *SLIM* excluding the regularization over the model parameters was introduced which achieved correct network sparsity (as shown in supplementary). Assuming priors over the model parameters created a bias over the generated network when compared to the ground truth network statistics.

Acknowledgements

We would like to express sincere appreciation and thank the reviewers for their constructive feedback and their insightful comments. We gratefully acknowledge the Independent Research Fund Denmark for supporting this work [grant number: 0136-00315B].

References

F. Atay and H. Tunçel Gölpek. On the spectrum of the normalized laplacian for signed graphs: Interlacing, contraction, and replication. *Linear Algebra and its Applications*, 442:165–177, 02 2014. doi: 10.1016/j.laa.2013.08.022.

P. Barberá, J. T. Jost, J. Nagler, J. A. Tucker, and R. Bonneau. Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological science*, 26(10):1531–1542, 2015.

O. Barndorff-Nielsen, D. Pollard, and N. Shephard. Integer-valued lévy processes and low latency financial econometrics. *Quantitative Finance*, 12, 10 2010.

S. Beugelsdijk, H. van Herk, and R. Maseland. The nature of societal conflict in europe; an archetypal analysis of the postmodern cosmopolitan, rural traditionalist and urban precariat. *JCMS*, n/a(n/a), 2022.

B. Büeler, A. Enge, and K. Fukuda. Exact volume computation for polytopes: a practical study. In *Polytopes—combinatorics and computation*, pages 131–154. Springer, 2000.

D. Cartwright and F. Harary. Structural balance: a generalization of heider's theory. *Psychological review*, 63 5: 277–93, 1956.

A. Çelikkanat, N. Nakis, and M. Mørup. Piecewise-velocity model for learning continuous-time dynamic node representations. In *The First Learning on Graphs Conference*, 2022.

K.-Y. Chiang, N. Natarajan, A. Tewari, and I. S. Dhillon. Exploiting longer cycles for link prediction in signed networks. In *CIKM*, page 1157–1162. Association for Computing Machinery, 2011.

A. Cutler and L. Breiman. Archetypal analysis. *Technometrics*, 36(4):338–347, 1994.

A. Dagnes. Us vs. them: Political polarization and the politicization of everything. In *Super Mad at Everything All the Time*, pages 119–165. Springer, 2019.

D. DellaPosta, Y. Shi, and M. Macy. Why do liberals drink lattes? *American Journal of Sociology*, 120(5):1473–1511, 2015.

V. R. K. Garimella and I. Weber. A long-term analysis of polarization on twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, 11: 528–531, May 2017.

G. H. Golub and C. F. Van Loan. *Matrix Computations (3rd Ed.)*. Johns Hopkins University Press, USA, 1996. ISBN 0801854148.

A. Grover and J. Leskovec. Node2Vec: Scalable feature learning for networks. In *KDD*, pages 855–864, 2016.

P. Guerra, W. Meira Jr, C. Cardie, and R. Kleinberg. A measure of polarization on social media networks based on community boundaries. In *Proceedings of the international AAAI conference on web and social media*, volume 7, pages 215–224, 2013.

F. Harary. On the notion of balance of a signed graph. *Michigan Mathematical Journal*, 2(2):143 – 146, 1953.

Y. Hart, H. Sheftel, J. Hausser, P. Szekely, N. B. Ben-Moshe, Y. Korem, A. Tendler, A. E. Mayo, and U. Alon. Inferring biological tasks using pareto analysis of high-dimensional data. *Nature methods*, 12(3): 233–235, 2015.

M. J. Hetherington. Review article: Putting polarization in perspective. *British Journal of Political Science*, 39 (2):413–448, 2009.

P. D. Hoff. Bilinear mixed-effects models for dyadic data. *JASA*, 100(469):286–295, 2005.

P. D. Hoff, A. E. Raftery, and M. S. Handcock. Latent space approaches to social network analysis. *JASA*, 97(460): 1090–1098, 2002.

J. Huang, H. Shen, L. Hou, and X. Cheng. Signed graph attention networks, 2019. URL `https://arxiv.org/abs/1906.10958`.

J. Huang, H. Shen, L. Hou, and X. Cheng. Sdgnn: Learning node representation for signed directed networks, 2021. URL `https://arxiv.org/abs/2101.02390`.

Z. Huang, A. Silva, and A. Singh. Pole: Polarized embedding for signed networks. *WSDM*, pages 390–400, 2022.

M. R. Islam, B. Aditya Prakash, and N. Ramakrishnan. SIGNet: Scalable embeddings for signed networks. In D. Phung, V. S. Tseng, G. I. Webb, B. Ho, M. Ganji, and L. Rashidi, editors, *Advances in Knowledge Discovery and Data Mining*, pages 157–169, Cham, 2018. Springer International Publishing.

D. Karlis and I. Ntzoufras. Bayesian analysis of the differences of count data. *Statistics in medicine*, 25:1885–905, 06 2006.

D. Karlis and I. Ntzoufras. Bayesian modelling of football outcomes: using the Skellam's distribution for the goal difference. *IMA Journal of Management Mathematics*, 20(2):133–145, 2008.

A. Keuchenius, P. Törnberg, and J. Uitermark. Why it is important to consider negative ties when studying polarized debates: A signed network analysis of a dutch cultural controversy on twitter. *PLOS ONE*, 2021.

B. Kim, K. H. Lee, L. Xue, and X. Niu. A review of dynamic network models with latent variables. *Statistics surveys*, 12:105, 2018a.

J. Kim, H. Park, J.-E. Lee, and U. Kang. SIDE: Representation learning in signed directed networks. In *Proceedings of the 2018 World Wide Web Conference*, page 509–518. International World Wide Web Conferences Steering Committee, 2018b.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization, 2017.

S. Kumar, W. L. Hamilton, J. Leskovec, and D. Jurafsky. Community interaction and conflict on the web. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, pages 933–943. International World Wide Web Conferences Steering Committee, 2018.

J. Leskovec, D. Huttenlocher, and J. Kleinberg. Predicting positive and negative links in online social networks. In *WWW*, page 641–650, 2010.

J. Moody and P. J. Mucha. Portrait of political party polarization. *Network Science*, 1(1):119–121, 2013. doi: 10.1017/nws.2012.3.

M. Mørup and L. Kai Hansen. Archetypal analysis for machine learning. In *2010 IEEE International Workshop on Machine Learning for Signal Processing*, pages 172–177, 2010. doi: 10.1109/MLSP.2010.5589222.

N. Nakis, A. Çelikkanat, S. L. Jørgensen, and M. Mørup. A hierarchical block distance model for ultra low-dimensional graph representations. 2022. URL `https://arxiv.org/abs/2204.05885`.

N. Nakis, A. Çelikkanat, and M. Mørup. HM-LDM: A hybrid-membership latent distance model. In *Complex Networks and Their Applications XI*, pages 350–363, Cham, 2023. Springer International Publishing. ISBN 978-3-031-21127-0.

Z. P. Neal. A sign of the times? weak and strong polarization in the u.s. congress, 1973–2016. *Social Networks*, 60:103–112, 2020.

B. Ordozgoiti, A. Matakos, and A. Gionis. Finding large balanced subgraphs in signed networks. In *Proceedings of The Web Conference 2020*, page 1378–1388, 2020.

B. Perozzi, R. Al-Rfou, and S. Skiena. DeepWalk: Online learning of social representations. *CoRR*, abs/1403.6652, 2014.

A. E. Raftery, X. Niu, P. D. Hoff, and K. Y. Yeung. Fast inference for the latent space network model using a case-control approximate likelihood. *Journal of Computational and Graphical Statistics*, 21(4):901–919, 2012.

P. Sarkar and A. Moore. Dynamic social network analysis using latent space models. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *NeurIPS*, volume 18, 2005.

O. Shoval, H. Sheftel, G. Shinar, Y. Hart, O. Ramote, A. Mayo, E. Dekel, K. Kavanagh, and U. Alon. Evolutionary trade-offs, pareto optimality, and the geometry of phenotype space. *Science*, 336(6085):1157–1160, 2012.

J. G. Skellam. The frequency distribution of the difference between two poisson variates belonging to different populations. *Journal of the Royal Statistical Society. Series A (General)*, 109(Pt 3):296–296, 1946.

M. Thomas, B. Pang, and L. Lee. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. *CoRR*, abs/cs/0607062, 2006.

R.-C. Tzeng, B. Ordozgoiti, and A. Gionis. Discovering conflicting groups in signed networks. In *NeurIPS*, 2020.

S. Wang, J. Tang, C. Aggarwal, Y. Chang, and H. Liu. *Signed Network Embedding in Social Media*, pages 327–335. 2017.

R. West, H. S. Paskov, J. Leskovec, and C. Potts. Exploiting social network structure for person-to-person sentiment analysis. *TACL*, 2:297–310, 2014.

E. Williamson. Plato's "eidos" and the archetypes of jung and frye. *Interpretations*, 16(1):94–104, 1985. ISSN 0196903X.

P. Xu, J. Wu, W. Hu, and B. Du. Link prediction with signed latent factors in signed social networks. *Proceedings of the Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, pages 1046–1054, 2019.

L. Zhuang, C.-H. Lin, M. A. Figueiredo, and J. M. Bioucas-Dias. Regularization parameter selection in minimum volume hyperspectral unmixing. *IEEE Transactions on Geoscience and Remote Sensing*, 57(12): 9858–9877, 2019.

For the published version please visit Advances in Complex Systems: https://doi.org/10.1142/S0219525923400027

A Hybrid Membership Latent Distance Model for Unsigned and Signed Integer Weighted Networks

Nikolaos Nakis
Department of Applied Mathematics and Computer Science, Technical University of Denmark, Anker Engelunds Vej 101 Kongens Lyngby, 2800, Denmark
nnak@dtu.dk

Abdulkadir Çelikkanat
Department of Applied Mathematics and Computer Science, Technical University of Denmark, Anker Engelunds Vej 101 Kongens Lyngby, 2800, Denmark
abce@dtu.dk

Morten Mørup
Department of Applied Mathematics and Computer Science, Technical University of Denmark, Anker Engelunds Vej 101 Kongens Lyngby, 2800, Denmark
mmor@dtu.dk

Graph representation learning (GRL) has become a prominent tool for furthering the understanding of complex networks providing tools for network embedding, link prediction, and node classification. In this paper, we propose the Hybrid Membership-Latent Distance Model (HM-LDM) by exploring how a Latent Distance Model (LDM) can be constrained to a latent simplex. By controlling the edge lengths of the corners of the simplex, the volume of the latent space can be systematically controlled. Thereby communities are revealed as the space becomes more constrained, with hard memberships being recovered as the simplex volume goes to zero. We further explore a recent likelihood formulation for signed networks utilizing the Skellam distribution to account for signed weighted networks and extend the HM-LDM to the signed Hybrid Membership-Latent Distance Model (sHM-LDM). Importantly, the induced likelihood function explicitly attracts nodes with positive links and deters nodes from having negative interactions. We demonstrate the utility of HM-LDM and sHM-LDM on several real networks. We find that the procedures successfully identify prominent distinct structures, as well as how nodes relate to the extracted aspects providing favorable performances in terms of link prediction when compared to prominent baselines. Furthermore, the learned soft memberships enable easily interpretable network visualizations highlighting distinct patterns.

Keywords: Signed Networks; Community Detection; Non-negative Matrix Factorization; Graph Representation Learning; Latent Space Modeling;

1. Introduction

In various scientific disciplines, including but not limited to physics, sociology, science-of-science, and biology, networks naturally arise to describe different interactions. These contain spin glasses, friendship interactions, scholarly collaborations, protein-to-protein interactions, structural and functional brain connectivity, and many more [50]. In order to study these networks and understand their underlying structures, scientists turn to graph analysis tools. The most prominent way for analyzing networks lies in Graph Representation Learning (GRL) [67], which includes approaches capable of performing downstream tasks such as link prediction, node classification, network reconstruction, and community detection with superior performance when compared to prior classical methods. Contrary to GRL, traditional algorithms are characterized by limited flexibility and capacity since they utilize node and graph-level statistics requiring careful design of heuristics and usually high time complexity [13]. The main goal of GRL is to find a mapping, through a learning process, projecting a network into a low-dimensional (usually Euclidean) latent space where node similarity in the graph is translated to node similarity in the latent space, i.e., by positioning related nodes close in proximity in the latent space [15].

Early GRL approaches capitalized on Natural Language Processing (NLP) where they performed random walks to generate node sequences that correspond to sentences in terms of the NLP terminology [8, 12, 54, 56, 61]. The core idea lies in simulating random walks over graphs and optimizing the co-occurrence probability for node pairs based on their obtained distance through the simulated walks. Relatively recent pioneering works [14] have extended GRL to the deep learning theory, giving rise to Graph Neural Networks (GNN). Essentially, GNNs perform iterative message-passing extending convolution operations to graphs. One of their limitations is usually the need for node features or else meta-data to avoid the over-smoothing pitfall hampering performance [32] when the GNN model defines deep architectures. Another major category of approaches for GRL relies on matrix decomposition tools [55,56]. Such models learn representations based on the decomposition of a target matrix, which can be constructed to convey first and high-order nodal proximity information [53, 56]. Despite Non-negative Matrix Factorization (NMF) being a prevalent technique for unsupervised signal decomposition and approximation of multivariate non-negative data, few GRL methods utilize such a decomposition. NMF methods have attracted considerable interest since they can extract interpretable part-based representations by revealing the latent factors of the imposed decomposition, which aids in structure retrieval [36].

NMF has been utilized in the context of network analysis and GRL [3,41,64,66], enabling efficient, unsupervised, and overlapping community detection. This has been explored in various studies, including a mixed-membership stochastic block model (MM-SBM) [1] defined based on a symmetric-NMF decomposition [41]. This method allows for part-based community assignments for networks while provid-

ing uniqueness guarantees. To obtain the propensity of nodes belonging to different communities, standard least-squares NMF optimization was replaced with a Poisson likelihood optimization [3]. Another study used a Poisson distribution to infer mixed memberships for overlapping community detection [64]. These studies involve the generation of mixed-membership vectors for part-based representations. These vectors are then projected onto a space generated by an NMF method, which captures abstract representations of node similarities, positions, and metric properties. Another popular application for NMF is the hyperspectral unmixing [6] via variational minimum volume regularization [17, 69]. A well-known approach is the Minimum Volume Constrained-Nonnegative Matrix Factorization (MVC-NMF) [42], which tries to approximate the hyperspectral data matrix with minimum error while including a volume constraint on the simplex matrix. MVC-NMF uses an alternating minimization procedure alternating over a quadratic programming problem and a nonconvex programming problem.

The Latent Space Models (LSMs) are also one of the most powerful ways to learn low-dimensional latent representations [49, 70]. These methods employ generalized linear models for constructing latent node embeddings which express important network characteristics. More specifically, the LDM [20] utilizes the Euclidean norm for positioning similar nodes closer in the latent space, which comes as a direct consequence of the triangular inequality, naturally representing transitivity (*"a friend of a friend is a friend"*) and homophily (*a tendency where similar nodes are more likely to connect to each other than dissimilar ones*) properties. An immediate result of obeying the triangular inequality is that the LDM successfully models high-order interactions, as present in complex systems [4, 44]. The LDM can be generalized through the Eigenmodel [19] that can account for stochastic equivalence (*"groups of nodes defined by shared intra- and inter-group relationships"*) akin to the SBM [1] and the mixed membership SBM [1]. Furthermore, LDMs have been endowed with a clustering model imposing a Gaussian Mixture Model as prior forming the latent position clustering model [16, 58].

Archetypal Analysis (AA) [10] has become a popular tool for extracting polytopes in tabular data. AA was originally defined as an unsupervised learning approach where input data are expressed as linear mixtures (convex combinations) of archetypes/distinct aspects being present in the data [45]. AA has been recently extended to the context of network analysis and the modeling of signed networks [48], characterizing polarization and conflict over graphs.

Unlike traditional networks modeling only positive and neutral links between entities, signed networks can capture more complex relations, such as cooperative and antagonistic approaches. They are instrumental in modeling more realistic and richer representations of real social structures. Hence, the analysis of the signed networks can reveal significant insights into understanding how the network structure is actually formed. The proverb *"The enemy of my enemy is my friend"* is a very known example demonstrating that driving forces leading individuals to form connections are not merely positive inclinations. The *balance theory* [25] explains

these motives by proposing that individuals have an inner desire to provide balance and consistency in their relationships. Inspired by the theory, POLE [24] proposes a novel network embedding method for signed networks based on generating random walks. It assigns a sign for each random walk by incorporating the balance theory. SIDE [30] also utilizes fixed-length random walks to extract the node representations of signed networks, but it employs a different optimization strategy. SIGAT [21], and SDGNN [22] propose approaches leveraging the successful graph neural network architectures for signed networks. The SLF approach [65] relies on extracting multiple latent factors to model four relationship types: positive, negative, non-link, and neutral. Most recently, SLDM [48] combined the latent space models and archetypal analysis to learn node embeddings reflecting the different aspects of networks, such as polarized groups or overlapping community structures.

This paper serves as an extension to the *HM-LDM: A Hybrid-Membership Latent Distance Model* paper as appeared in [47]. The main contributions of the paper and its extended version can be summarized as:

- We introduce a novel method for unsupervised representation learning on graphs called Hybrid-Membership Latent Distance Model (HM-LDM), which combines the strengths of LDM and NMF. The HM-LDM approach reconciles part-based network representations with low-dimensional latent spaces that satisfy similarity properties like homophily and transitivity. These properties play a critical role in GRL because they enable a straightforward interpretation of network structure. Moreover, the proposed method captures the latent community structure of the networks using a simple continuous optimization procedure based on the log-likelihood of the network. Unlike most existing methods that impose hard constraints on community memberships, the assignment of community memberships in our hybrid model can be controlled and altered using the simplex volume as defined by the latent node representations. We extensively evaluate the proposed method's performance in link prediction and community discovery tasks across various network types and demonstrate its superiority over existing methods.
- We hereby, extend the framework to the analysis of signed networks via the use of the Skellam distribution forming the signed Hybrid-Membership Latent Distance Model (sHM-LDM) inspired by recent advances in GRL [48]. The model characterizes and uncovers distinct aspects of signed networks by constraining the latent space to the D-simplex. We show that the sHM-LDM relates to archetypal analysis for relational data [48] as a minimal volume approach and as a special case when polytopes are constrained to simplexes. We benchmark the performance of our model against prominent signed network representation learning approaches and across four real signed networks, as well as two real bipartite networks.

Source code: *https://github.com/Nicknakis/HM-LDM*.

2. Problem statement and proposed method

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be an undirected graph where $\mathcal{V}$ shows the vertex set and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ the edge set. We use $\mathbf{Y}_{N\times N} = (y_{i,j})$ to denote the adjacency matrix of the graph where $y_{i,j} = 0$ if the pair $(i, j) \notin \mathcal{E}$ otherwise it is non-zero value for all $1 \leq i < j \leq N := |\mathcal{V}|$. It is worth noting that we will also consider signed weighted networks in the paper, so the edge weight or the entries of the adjacency matrix can take any positive or negative integer value ($y_{ij} \in \mathbb{Z}$). In the latter case, we will further denote $\mathcal{E}^+$ as the positive edge set, and $\mathcal{E}^-$ as the negative edge set. The detailed list of the symbols used throughout the manuscript and their corresponding definitions can be found in Table 1.

Our main goal is to learn a representation, $\mathbf{w}_i \in \mathbb{R}^D$, for each node $i \in \mathcal{V}$ in a lower dimensional space ($D \ll N$) such that similar nodes in the network should have close embeddings. More specifically, we concentrate on mapping the nodes into the unit D-simplex, $\Delta^D \subset \mathbb{R}_+^{D+1}$, which is defined by

$$\Delta^D = \left\{ (x_0, \ldots, x_D) \in \mathbb{R}^{D+1} \,\middle|\, \sum_{d=0}^{D} x_d = 1,\ x_d \geq 0,\ \forall d \in \{0, \ldots, D\} \right\}.$$

Consequently, for unsigned networks, the inferred node representations carry information about latent community memberships. While in the case of signed networks, node embeddings define memberships over distinct aspects and profiles being present in the network. Importantly, in contrast with other GRL approaches, in this study, we seek and construct identifiable solutions which can only be achieved up to a permutation invariance, as reported in Def. 1. Identifiability guarantees are also extended to the modeling of signed networks providing embedding spaces that can

Table 1: Table of symbols

Symbol	Description
$\mathcal{G}$	Graph
$\mathcal{V}$	Vertex set
$\mathcal{E}$	Edge set
$\mathcal{E}^+$	Positive edge set
$\mathcal{E}^-$	Negative edge set
N	Number of nodes
D	Dimension size
$\gamma_i, \beta_i, \psi_i$	Bias terms of node i
$\mathbf{w}_i$	Latent embedding for node i
λ_{ij}	Poisson rate (intensity) of node pair (i, j)
λ_{ij}^+	Positive interaction Poisson rate (intensity) of node pair (i, j) of the Skellam distribution
λ_{ij}^-	Negative interaction Poisson rate (intensity) of node pair (i, j) of the Skellam distribution
$\mathcal{I}_{\lvert y\rvert}$	Modified Bessel function of the first kind and order $\lvert y\rvert$
δ	Simplex side length with $\delta \in \mathbb{R}_+$
p	Power of the ℓ_2 norm with $p \in \{1, 2\}$
Δ^D	The standard $D-$simplex
$\mathbf{\Lambda}$	Eigenmodel non-negative relational matrix
$\mathbf{A}$	The matrix containing the archetypes (extreme points of the convex hull) with $\mathbf{A} \in \mathbb{R}^{(D+1)\times(D+1)}$

easily be interpreted.

In the following part, we will first introduce the Hybrid-Membership Latent Distance Model (HM-LDM) focused on unsigned networks, and later we will generalize the framework to the analysis of signed networks forming the Signed Hybrid-Membership Latent Distance Model (sHM-LDM).

Definition 1. (Identifiability) *Let* $\mathbf{W}$ *be an optimal embedding matrix whose rows indicate the corresponding node representations. We call* $\mathbf{W}$ *an identifiable solution up to a permutation if there is a matrix* $\mathbf{P}$ *satisfying* $\widetilde{\mathbf{W}} = \mathbf{WP}$ *for some optimal solution* $\widetilde{\mathbf{W}}$*, then* $\mathbf{P}$ *must be a permutation matrix.*

2.1. *The Hybrid-Membership Latent Distance Model*

For a given unsigned network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, we suppose that the random variables representing the links for a pair of nodes i and j independently follow a Poisson distribution when conditioned on the latent representations $\{\boldsymbol{W}, \boldsymbol{\gamma}\}$, as introduced later. In this section, we consider unweighted networks, so the entries of the adjacency matrix, $\mathbf{Y} = (y_{ij}) \in \{0,1\}^{|\mathcal{V}|\times|\mathcal{V}|}$ are binary values, and we can write the log-likelihood function as follows:

$$\log P(\mathbf{Y}|\boldsymbol{W}, \boldsymbol{\gamma}) = \sum_{\substack{i<j \\ y_{ij}=1}} \log(\lambda_{ij}(\mathbf{w}_i, \mathbf{w}_j, \gamma_i, \gamma_j)) \; - \sum_{i<j} \Big(\lambda_{ij}(\mathbf{w}_i, \mathbf{w}_j, \gamma_i, \gamma_j) + \log(y_{ij}!)\Big). \tag{1}$$

Similar to the work in [18], we here employ the Poisson regression approach for unweighted networks since it successfully generalizes to the modeling of binary networks [64].

We utilize the rates of the distributions to learn the representations of nodes in the latent space by defining the Poisson rate λ_{ij} as follows:

$$\log \lambda_{ij} = \Big(\gamma_i + \gamma_j - \delta^p \cdot ||\mathbf{w}_i - \mathbf{w}_j||_2^p\Big), \tag{2}$$

where $\mathbf{w_i} \in [0,1]^{D+1}$ are the latent embeddings constrained to the $D-$simplex, i.e. $\sum_{d=1}^{D+1} w_{id} = 1$, $\delta \in \mathbb{R}_+$ is the non-negative value controlling the simplex volume, and $\gamma_i \in \mathbb{R}$ a bias term of node $i \in \mathcal{V}$ accounting for node-specific effects [18, 33] such as degree heterogeneity. Lastly, p is the power of the ℓ_2 norm with $p \in \{1, 2\}$ controlling the model specification. Specifically, power p adjusts the influence of the embedding distances in the rate functions. In other words, in Eq. 2 we constrain the latent space to the $D-$simplex, and the simplex's edge lengths (1-faces) are scaled by the non-negative constant δ, controlling the simplex side length and thus the simplex volume. In the rest of the paper, we will call this proposed method by Hybrid-Membership Latent Distance Model (HM-LDM).

It can be seen that a non-negative Eigenmodel with bias terms (i.e. $\tilde{\gamma}_i + \tilde{\gamma}_j + (\tilde{\mathbf{w}}_i \boldsymbol{\Lambda} \tilde{\mathbf{w}}_j^\top)$) corresponds to Eq. (2) for $p = 2$ if $\boldsymbol{\Lambda}$ is chosen as a diagonal matrix with constant entries $2\delta^2$, and if the bias terms are reparameterized as $\tilde{\gamma}_i = \gamma_i - \delta^2 \cdot ||\mathbf{w}_i||_2^2$

since expression $\tilde{\gamma}_i + \tilde{\gamma}_j + (\tilde{\mathbf{w}}_i \mathbf{\Lambda} \tilde{\mathbf{w}}_j^\top)$ turns into:

$$(\gamma_i - \delta^2 ||\mathbf{w}_i||_2^2) + (\gamma_j - \delta^2 ||\mathbf{w}_j||_2^2) + (2\delta^2 \mathbf{w}_i \mathbf{w}_j^\top) = \gamma_i + \gamma_j - \delta^2 \|\mathbf{w}_i - \mathbf{w}_j\|^2.$$

Therefore, the squared Euclidean distance incorporates the conventional LDM to the non-negativity-constrained Eigenmodel. Although the squared Euclidean distance is not a metric, it still embodies the homophily property, resulting in an interpretable latent space. Despite not exactly satisfying the triangle inequality, it preserves the relative ordering of pairwise Euclidean distances. That's why it is highly preferred in many applications since it is a strictly convex smooth function. By using the well-known cosine formula, we can write:

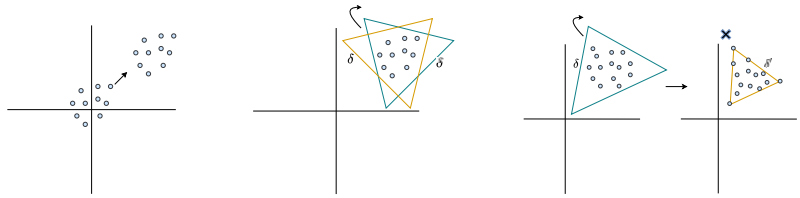
$$||\mathbf{w}_i - \mathbf{w}_j||_2^2 = ||\mathbf{w}_i - \mathbf{w}_k||_2^2 + ||\mathbf{w}_k - \mathbf{w}_j||_2^2 - 2||\mathbf{w}_i - \mathbf{w}_k||_2 ||\mathbf{w}_k - \mathbf{w}_j||_2 \cos(\theta),$$

where $\theta \in (-\pi/2, \pi/2)$ represents the angle between $\mathbf{w}_i - \mathbf{w}_k$ and $\mathbf{w}_k - \mathbf{w}_j$. Note that the third term also approaches to 0 for $\theta \to \pi/2$. For the case where $\theta \in [\pi/2, 3\pi/2]$, it satisfies the triangle inequality: $||\mathbf{w}_i - \mathbf{w}_j||_2^2 \leq ||\mathbf{w}_i - \mathbf{w}_k||_2^2 + ||\mathbf{w}_k - \mathbf{w}_j||_2^2$.

Since we learn the node representations in a $D-$simplex space, each entry of an embedding vector, in fact, points out a latent community membership, so the node representations also provide information regarding the community structure of the network. Note that we can translate the learned embeddings to the non-negative orthant without any loss in performance or in expressive capability since the translation is invariant to the distance metric, as shown in Fig 1 (a). A rotation operation also does not affect the pairwise distance among the embedding vectors but the node representations must be positioned inside a ball lying in a $D-$simplex otherwise, the embeddings cannot be rotated (see Fig 1 (b)).

However, as we mentioned before, the embedding vectors also define the nodes' community memberships. Therefore, a rotation operation alters the community assignments while leaving the distance matrix invariant. As a result, the latent representations cannot be used to express community information in this case. It is worth noticing that we can have *identifiable* node representations if the corners of the simplex include at least one node because it makes the rotation operation inapplicable. In this regard, this condition can be satisfied by the distance scaling parameter (i.e., $\delta \in \mathbb{R}^+$) introduced in Eq. 2. Since we know that $\|x\|_1^p / \sqrt{D^p} \leq \|x\|_2^p \leq \|x\|_1^p$ for $p \in \{1, 2\}$, shrinking the volume of the simplex sufficiently (equivalently decreasing the δ value) forces nodes to populate around the corners of the simplex. The node embeddings move towards the corners of the simplex to balance the change in the term $\delta^p \|\mathbf{w}_i - \mathbf{w}_j\|_2^p$ since we have $\|\mathbf{w}_i\| = 1$ for all $i \in \mathcal{V}$.

We will name a node *champion* if it is located in one of the corners of the simplex. In other words, its latent representation must be a standard binary unit vector in a $D-$simplex space. The champion nodes play a crucial role in achieving identifiability since the learned representations become identifiable (up to a permutation matrix) if every corner of the simplex is occupied by at least one champion node (please see the definition below). In this case, any random rotation will no longer leave the solution invariant, as illustrated in Fig 1 (c). Hence, the scaling parameter,



(a) Translation invariances.

(b) Rotation invariances.

(c) Decreased simplex volume ensuring identifiability.

Fig. 1: A 2-dimensional latent space with the 2-simplex given as the green and yellow triangles, the blue points denote embedding positions of the LDM and δ is the simplex size.

δ, determines the model's type of memberships and expressive capabilities. Large values of δ make the solution rotationally invariant. On the other hand, small values of δ result in identifiable solutions and hard cluster assignments, where nodes are exclusively assigned to the corners of the simplex. Moreover, certain regimes of δ values can provide identifiable solutions with similar performance to LDM.

Definition 2. (Community champion) *A node for a latent community is called champion if it belongs to the community (simplex corner) while forming a binary unit vector.*

We can approach model identifiability for the $p = 2$ model specification from a different perspective using the Non-negative Matrix Factorization (NMF) theory. We achieve this by re-parameterizing Eq. (2) with $\tilde{\gamma}_i + \tilde{\gamma}_j + 2\delta^2 \cdot (\mathbf{w}_i \mathbf{w}_j^\top)$ as previously discussed. In this formulation, the product $\mathbf{W}\mathbf{W}^\top$ defines a symmetric NMF problem that is uniquely factorized (up to permutation invariance) and identifiable when $\mathbf{W}$ is full-rank, and at least one node resides solely in each corner of the simplex, ensuring separability condition [23, 41]. Under this NMF formulation, the product $\mathbf{w}_i \mathbf{w}_j^\top \in [0, 1]$ reaches its upper bound only if nodes i and j reside in the same corner of the simplex. When δ is small, the model favors hard latent community assignments of nodes since nodes with similar features achieve high values only when they belong to the same latent community (simplex corner). On the other hand, when nodes head towards the corners of the simplex for large values of δ, the second term of the log-likelihood function in Eq. (1) changes exponentially. Hence, assigning dissimilar nodes to the same community severely penalizes the likelihood. For this reason, a high value of δ is beneficial for mixed-membership allocations.

2.2. *The Signed Hybrid-Membership Latent Distance Model*

Recent advances in GRL [48], extended LDMs to the study of signed networks while characterizing network polarization via the use of Archetypal Analysis (AA) [10, 45]

and the Skellam distribution [59]. The Skellam distribution is the difference of two independent Poisson-distributed random variables ($y = N_1 - N_2 \in \mathbb{Z}$) with respect to the rates λ^+ and λ^-:

$$P(y|\lambda^+, \lambda^-) = e^{-(\lambda^+ + \lambda^-)} \left(\frac{\lambda^+}{\lambda^-}\right)^{y/2} \mathcal{I}_{|y|}\left(2\sqrt{\lambda^+\lambda^-}\right), \tag{3}$$

where $N_1 \sim Pois(\lambda^+)$ and $N_2 \sim Pois(\lambda^-)$, and $\mathcal{I}_{|y|}$ is the modified Bessel function of the first kind and order $|y|$.

Whereas in [48] the network representations were constrained to the convex hull as defined by the inferred representations, it is discussed that other approaches to model pure/ideal forms have been Minimal Volume (MV) approaches as defined by

$$\mathbf{Z} \approx \mathbf{AW} \quad \text{s.t. } vol(\mathbf{A}) = v \text{ and } \boldsymbol{w}_j \in \Delta^D, \tag{4}$$

where $\mathbf{A} \in \mathbb{R}^{(D+1)\times(D+1)}$ is the matrix describing the archetypes (extreme points of the convex hull) of the latent space, and $vol(\mathbf{A})$ is the volume of matrix $\mathbf{A}$ which can be expressed through the determinant as $|det(\mathbf{A})|$, when $\mathbf{A}$ is a square matrix [17, 69]. Extraction of distinct aspects/profiles through MV does not require the presence of "pure" observations defining the convex-hull or else the extracted polytope/simplex. As the volume decreases, observations are "forced" to populate the corners of the polytope, yielding archetypal characterization when the reconstruction of data is defined through convex combinations of these corners.

The main disadvantage of MV procedures is the need for careful regularization tuning to define volumes ensuring identifiability as well as maintaining enough capacity to express the data with a small reconstruction error [69]. In addition, analytical and tractable computation of the volume of polytopes requires calculating the sum of determinants for all simplexes used to construct the inferred polytope [7]. This is computationally expensive (especially in high dimensions) and sometimes unstable when $\mathbf{A}$ comes close to singular.

In this paper, we constrain the columns of matrix $\mathbf{A}$ to the D-simplex with length δ. Thus, by controlling the volume of $\mathbf{A}$, we essentially define a constrained-to-simplexes MV approach. Calculating the volume for the $D-$simplex with length δ is straightforward and computationally efficient. Rather than including regularization over the volume of $\mathbf{A}$ in the loss function during inference, we deterministically control the simplex length δ which is given as an input to the model and gradually decreased until uniqueness guarantees are obtained. Volume minimization can be obtained trivially by decreasing δ. Such a procedure gives us explicit control over the model capacity by fixing the volume which is harder to be obtained with classical MV approaches where the volume expression is inserted in the loss function.

Essentially, by defining $\mathbf{A}$ as $\mathbf{A} = \delta \cdot \mathbf{I}$, with $\mathbf{I}$ being the $(D+1)\times(D+1)$ identity matrix, we obtain as a special case of archetypal analysis under a constrained MV formulation. In addition, if every corner of the introduced simplex is populated by at least one node champion we obtain unique representations defining hybrid memberships.

We now introduce the signed Hybrid-Membership Latent Distance Model (sHM-LDM). The sHM-LDM is able to analyse signed networks, and similar to [48] it introduces two Skellam rate parameters for Eq. (3) as:

$$\lambda_{ij}^{+} = \exp\left(\beta_i + \beta_j - \delta^p ||\mathbf{w}_i - \mathbf{w}_j||_2^p\right), \tag{5}$$

$$\lambda_{ij}^{-} = \exp\left(\psi_i + \psi_j + \delta^p ||\mathbf{w}_i - \mathbf{w}_j||_2^p\right), \tag{6}$$

where again $\mathbf{w_i} \in [0,1]^{D+1}$ and $\sum_{d=1}^{D+1} w_{id} = 1$, $\delta \in \mathbb{R}_+$ and $\beta_i, \psi_j \in \mathbb{R}$ denote the node-specific random-effects. As explained in [48], β_i, β_j represent the "social" effects/reach of a node and the tendency to form (as a receiver and as a sender, respectively) positive interactions, expressing positive degree heterogeneity (indicated by + as a superscript of λ). In contrast, ψ_i, ψ_j provides the "anti-social" effect/reach of a node to form negative connections and thus models negative degree heterogeneity (indicated by − as a superscript of λ). The norm degree $p \in \{1, 2\}$ controls the power of the ℓ^2-norm, and thus the model specification, as in the unsigned case.

As in [48], we define a maximum-a-posteriori (MAP) estimation, utilizing the Skellam likelihood over the adjacency matrix $\mathbf{Y}$ of the network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. We conditionally assume an independent likelihood given the unobserved latent positions and random effects. The corresponding loss function excluding constant terms is:

$$L = \sum_{i<j} \left(\lambda_{ij}^{+} + \lambda_{ij}^{-} - \frac{y_{ij}}{2} \log\left(\frac{\lambda_{ij}^{+}}{\lambda_{ij}^{-}}\right) \right) - \sum_{i<j} \log I_{|y_{ij}|}\left(2\sqrt{\lambda_{ij}^{+}\lambda_{ij}^{-}}\right) + \frac{\rho}{2}\left(||\boldsymbol{\beta}||_F^2 + ||\boldsymbol{\psi}||_F^2\right), \tag{7}$$

where $\mathcal{I}_{|y|}$ is the modified Bessel function of the first kind and order $|y|$, $||\cdot||_F$ denotes the Frobenius norm. In addition, ρ is the regularization strength where $\rho = 1$ is assumed throughout this paper yielding a normal prior with zero mean and unit variance for the random effects. For the latent positions, we assume a uniform Dirichlet distribution as a prior which only adds a constant term in Eq. 7 and thus is excluded.

Choosing the case where $p = 2$, meaning that the sHM-LDM utilizes the squared Euclidean norm, we are able once more to relate the model to an Eigenmodel by creating the following reparameterizations. For the rate responsible for positive interactions $\{\lambda_{ij}^{+}\}$ as: $\tilde{\beta}_i + \tilde{\beta}_j + (\tilde{\mathbf{w}}_i \boldsymbol{\Lambda} \tilde{\mathbf{w}}_j^\top)$ where $\boldsymbol{\Lambda}$ is a diagonal matrix having non-negative elements, i.e. $\tilde{\beta}_i = \beta_i - \delta^2 \cdot ||\mathbf{w}_i||_2^2$, $\tilde{\beta}_j = \beta_j - \delta^2 \cdot ||\mathbf{w}_j||_2^2$ and $\tilde{\mathbf{w}}_i \boldsymbol{\Lambda} \tilde{\mathbf{w}}_j^\top = 2\delta^2 \cdot \mathbf{w}_i \mathbf{w}_j^\top$. Similarly, for the rate responsible for expressing animosity $\{\lambda_{ij}^{-}\}$ as: $\tilde{\psi}_i + \tilde{\psi}_j + (\tilde{\mathbf{w}}_i \boldsymbol{\Lambda} \tilde{\mathbf{w}}_j^\top)$ where $\boldsymbol{\Lambda}$ is a diagonal matrix having non-positive elements, i.e. $\tilde{\psi}_i = \psi_i - \delta^2 \cdot ||\mathbf{w}_i||_2^2$, $\tilde{\psi}_j = \psi_j - \delta^2 \cdot ||\mathbf{w}_j||_2^2$ and $\tilde{\mathbf{w}}_i \boldsymbol{\Lambda} \tilde{\mathbf{w}}_j^\top = 2\delta^2 \cdot \mathbf{w}_i \mathbf{w}_j^\top$. We witness that homophily in the case of sHM-LDM is expressed through a non-negative Eigenmodel (as in the unsigned case) while animosity/heterophily is expressed through a non-positive Eigenmodel able to express stochastic equivalence [19]. These two formulations admit the same embedding matrix $\mathbf{W}$ which balances the expression of "opposing" forces (homophily and animosity) in the latent space. Lastly, for $p = 2$ both expressions admit to an NMF operation, obtaining

Table 2: Network statistics; $|\mathcal{V}|$: # Nodes, $|\mathcal{E}|$: # Edges, $|\mathcal{K}|$: # Communities.

	AstroPh [39]	*GrQc* [39]	*Facebook* [39]	*HepTh* [39]	*Hamilton* [43]	*Amherst* [43]	*Rochester* [43]	*Mich* [43]
$\lvert\mathcal{V}\rvert$	17,903	5,242	4,039	8,638	2,118	2,021	4,145	2,933
$\lvert\mathcal{E}\rvert$	197,031	14,496	88,234	24,827	87,486	87,496	145,305	54,903
$\lvert\mathcal{K}\rvert$	-	-	-	-	15	15	19	13

an identifiable and unique factorization (up to permutation invariance) when $\mathbf{W}$ is full-rank and at least one node resides solely in each simplex corner [23, 41] as in the case of HM-LDM for unsigned networks.

3. Experimental evaluation

We continue by assessing the effectiveness and efficacy of the suggested techniques. We start with the case of unsigned networks, including both latent and ground-truth community structures, and test HM-LDM based on its capability to detect identifiable latent structures as well as to perform link prediction. Additionally, for the networks with known community structures, we assess how the model can successfully infer the ground-truth community labels. We then continue with the case of signed networks for evaluating the performance of sHM-LDM in its ability to perform signed link prediction and discovery of distinct profiles.

For both the training of HM-LDM and sHM-LDM, we make use of the Adam optimizer [31], minimizing for the two models the log-likelihood function of Eq. (1) and the MAP expression of Eq. (7), respectively. The learning rate is set as $lr \in [0.01, 0.1]$. The node-specific random effects vectors for all models are randomly initialized and then tuned separately (for 1000 iterations) by detaching initially the gradients from the latent representations $\mathbf{W}$. The latent embeddings matrix $\mathbf{W}$ is initialized based on the eigenvalues obtained by the spectral decomposition of the normalized Laplacian matrix of the network as expressed for unsigned [27, 51] and signed [2] networks.

3.1. *Unsigned Network Experiments*

We consider eight unsigned networks of various sizes and structures. We hereby supply the reader with additional information for the considered networks. The four networks with unknown community labels include (i) *AstroPh*, (ii) *GrQc*, and (iii) *HepTh* [38] are collaboration networks based on papers submitted to the astrophysics, general relativity and quantum cosmology, and high energy physics categories of the e-print ArXiv, respectively. An edge between a pair of nodes (representing authors) is created if they have co-authored a paper. (iv) *Facebook* [40] is a social network based on data obtained by a survey on a Facebook application. The additional four networks with ground-truth community labels include (v) *Hamilton*, (vi) *Amherst*, (vii) *Rochester*, and (viii) *Mich* which are all Facebook networks describing online friendships/connections of four American universities with the class

year serving as the ground truth community [43]. Network statistics are summarized by Table 2. We treat the above networks as unweighted and undirected.

For the experiments, we consider eleven various prominent graph representation learning methods to evaluate the performance of our proposed approach. These are: (i) DeepWalk [54], (ii) Node2Vec [12] which are two random-walk based methods. (iii) LINE [61] learning node embeddings vectors by optimizing the first- and second-order proximity information. (iv) NetMF [56] that factorizes the pointwise mutual information matrix of node co-occurrences obtained by random walks. (v) NetSMF [55] the scalable extension of the NetMF method [56]. (vi) LouvainNE [5] obtaining node representations by aggregating hierarchical embeddings of extracted network sub-graphs. (vii) ProNE [68] which finds representation based on a sparse matrix factorization and spectral propagation operations. We also consider four NMF-based embedding approaches able to convey information about community memberships. These include (viii) *NNSED* utilizing an encoder-decoder approach for community detection. (ix) *MNMF* unifying NMF representation learning with modularity-based community detection. (x) *BigClam* defining a model-based community detection algorithm able to detect overlapping community structures. (xi) *SymmNMF* decomposing a pairwise similarity measure matrix between nodes of the network admitting graph clustering properties.

Link prediction: To conduct the link prediction experiments, we adopt a commonly used approach [12, 54], where we eliminate half of the network edges while ensuring that the remaining network stays connected. The removed edges, together with the equivalent number of node pairs (that were not part of the original network edges), create the negative instances for the test set. The models learn network embeddings based on the remaining network.

For the link prediction experiments, we use the four networks with unknown community structures and compare the performance in Table 3, in terms of the Area Under Curve-Receiver Operating Characteristic (AUC-ROC) metric. We benchmark HM-LDM against other notable GRL and NMF models while considering the performance across various dimensions. All baselines are fine-tuned, and feature vectors for dyads are generated using binary operators (average, Hadamard, weighted-L1, weighted-L2) [12]. For the baselines, we further train a logistic regression model with L_2 regularization and based on the constructed feature vectors make link predictions. Specifically, we choose the optimal hyperparameters and binary operator for each baseline model, based on which operator and hyperparameters return the highest AUC-ROC score.

For our frameworks, we follow a different approach leading to an unbiased estimation of link prediction performance. More specifically, we report results based on the first δ value (as we decrease the volume) that makes the solution identifiable, meaning the δ value where at least one community champion resides in a simplex corner. Importantly, there exist additional values for δ which define identifiable solutions as well as increased performance with respect to the reported one

Table 3: Area Under Curve (AUC-ROC) scores for varying representation sizes.

	AstroPh			*GrQc*			*Facebook*			*HepTh*		
Dimension (D)	8	16	32	8	16	32	8	16	32	8	16	32
DeepWalk [54]	.945	.950	.952	.919	.916	.929	.986	.986	.984	.874	.867	.873
Node2Vec [12]	.950	.962	.957	.897	.913	.930	.988	.988	.987	.881	.882	.881
LINE [61]	.909	.938	.947	.920	.925	.919	.981	.987	.983	.873	.886	.882
NetMF [56]	.813	.823	.839	.860	.866	.877	.935	.963	.971	.792	.806	.821
NetSMF [55]	.891	.901	.919	.837	.858	.886	.975	.981	.985	.809	.822	.836
LouvainNE [5]	.813	.811	.819	.868	.875	.873	.958	.961	.963	.874	.867	.873
ProNE [68]	.907	.929	.947	.885	.911	.921	.971	.982	.987	.827	.846	.859
NNSED [60]	.861	.882	.891	.792	.808	.828	.908	.927	.935	.756	.779	.796
MNMF [62]	.893	.925	.943	.911	.928	.937	.965	.978	.982	.857	.880	.891
BigClam [66]	.500	.723	.810	.752	.769	.780	.744	.722	.647	.776	.700	.748
SymmNMF [34]	.767	.779	.800	.729	.772	.835	.933	.942	.951	.696	.727	.766
HM-LDM($p = 1$)	.956	.952	.952	**.944**	**.948**	**.951**	.982	.979	.974	**.916**	**.921**	**.924**
HM-LDM($p = 2$)	**.972**	**.973**	**.963**	.940	.942	.946	**.992**	**.993**	**.993**	.908	.910	.911

but are disregarded so the evaluation stays unbiased. In addition, predictions for HM-LDM are based directly on the Poisson rates λ_{ij} defined for test set pairs $\{i, j\}$ with AUC-ROC scores as reported in Table 3. This comes as an advantage of HM-LDM since it defines a likelihood function over the network connections and thus has no need for post-processing steps (such as training a logistic regression model) to make predictions. The true dimensions for HM-LDM are $D + 1$ but reported as D since this is the true number of model parameters, for a fair comparison with the baselines. Results for our method are reported based on the average performance over five independent runs of the model (error bars were found to be in the scale of 10^{-3} and thus not presented).

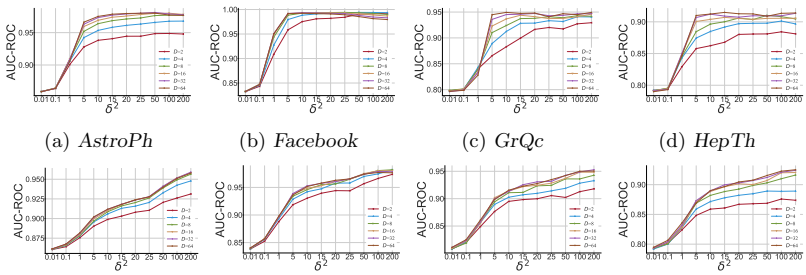
Upon contrasting our findings with the non-NMF models, we found that our HM-LDM (either $p = 1$ or $p = 2$) outperforms these baselines and, in most cases, by a significant margin, producing favorable results. We notice a considerable difference in performance when comparing HM-LDM with other part-based representation models, indicating the existence of identifiable regimes based on δ values where we can successfully obtain community memberships while simultaneously demonstrating the link prediction abilities of unconstrained LDM. (AUC Precision-Recall scores are similar to the AUC-ROC scores and thus not presented)

Performance and simplex sizes: Fig 2 displays the AUC-ROC scores in terms of link prediction performance as a function of δ^2 for various latent dimensions, and networks, and both $p = 1$ and $p = 2$. As expected, we here understand that small δ values provide the minimum scores. This is a direct consequence of the fact that homophily properties are not adequately met (except within clusters) due to the very small simplex volumes that these low δ values constrain the latent space to. If we think of HM-LDM with $p = 2$ as a positive Eigenmodel, we can also see how the positivity constraint on the Λ diagonal matrix hinders the expression of stochastic equivalence, which would boost performance even on low simplex volumes. As we increase δ values, we naturally approach the performance of an unconstrained LDM.

Comparing the case of p=2 (squared), and p=1 (simple) for the ℓ^2-norm, we observe that the former reaches performance saturation more rapidly.

Type and quality of latent memberships: We here study how the size of the simplex affects the membership types of HM-LDM. Fig 3 illustrates how the percentage of community champions (nodes assigned to simplex corners) for HM-LDM as a function of δ^2 and for different latent dimensions. When δ is small, almost all nodes are exclusively assigned to a simplex corner, resulting in hard assignments to clusters. As δ increases, more nodes are assigned with mixed memberships, while the number of champions decreases to zero for large δ values in all dimension cases. The decrease in community champions is steeper for $p = 2$ compared to $p = 1$. This also explains why the squared ℓ^2 choice leads to faster convergence in AUC-ROC, as the model converges faster to the classic LDM. It is evident that the $p = 2$ HM-LDM requires smaller simplex volumes to be identifiable. In Fig 4, we provide the reorganized adjacency matrices with community allocations given by HM-LDM, showing how the model successfully uncovers latent communities and produces part-based network representations while identifiability is ensured by choosing appropriate δ values, or equivalently appropriate simplex volumes. (for mixed-memberships nodes are assigned to the cluster in which they express the maximum membership)

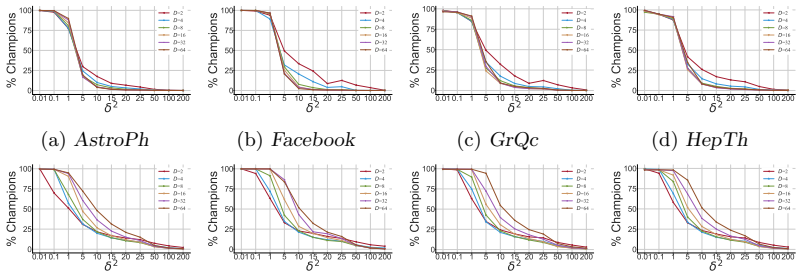
Experiments using real ground-truth communities: To evaluate the effectiveness of HM-LDM in identifying meaningful communities, we conduct experiments using four networks with known ground-truth community labels. For NMF-based methods, including our own, we assess the algorithms' ability to identify correct structures by comparing the inferred memberships with the ground-truth labels. We set the number of latent dimensions equal to the total number of



(a) *AstroPh* (b) *Facebook* (c) *GrQc* (d) *HepTh*

(e) *AstroPh* (f) *Facebook* (g) *GrQc* (h) *HepTh*

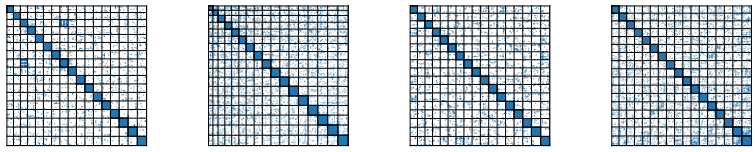
Fig. 2: AUC-ROC scores as a function of δ^2 (simplex size) across dimensions for HM-LDM. Increasing δ^2 (simplex volume) leads to higher performance as the model becomes more flexible until saturation (unconstrained LDM regime). Top row: $p = 2$ model specification. Bottom row $p = 1$ model specification.



(a) *AstroPh* (b) *Facebook* (c) *GrQc* (d) *HepTh*

(e) *AstroPh* (f) *Facebook* (g) *GrQc* (h) *HepTh*

Fig. 3: Total community champions (%) in terms of δ^2 (simplex size) across dimensions for HM-LDM. Decreasing δ^2 (simplex volumes) leads to a higher percentage of nodes positioned on the simplex corners (equivalent to hard clustering) until all nodes are pushed on the corners for very small volumes. Top row: $p = 2$ model specification. Bottom row $p = 1$ model specification.



(a) *GrQc* ($p = 2$) (b) *HepTh* ($p = 2$) (c) *GrQc* ($p = 1$) (d) *HepTh* ($p = 1$)

Fig. 4: Ordered adjacency matrices based on the memberships of a $D = 16$ dimensional HM-LDM with δ values ensuring identifiability, empirically showcasing community extraction and identification.

communities. For GRL approaches that do not provide memberships, we extract latent embeddings and use *k-means* to assign communities. We report the Normalized Mutual Information (NMI) score and Adjusted Rand Index (ARI), which are well-established measures for community quality assessment [9]. We tune all baseline methods separately for each network in terms of their hyperparameters. In contrast, for HM-LDM, we do not perform any tuning and just set $\delta = 1$ for all networks, resulting in informative and mostly hard cluster assignments. We report scores averaged over five independent runs of the Adam optimizer, each of which includes five additional runs, selecting the model with the lowest training loss to avoid the effect of local minimas. We summarize our findings in Table 4, where we witness a mostly favorable or on-par performance of HM-LDM with all of the

Table 4: Normalized Mutual Information (NMI) and Adjusted Rand Index (ARI) scores for networks with ground-truth communities.

	Amherst		*Rochester*		*Mich*		*Hamilton*	
Metric	NMI	ARI	NMI	ARI	NMI	ARI	NMI	ARI
DeepWalk [54]	.498	.347	.348	.205	.207	.157	.447	.303
Node2Vec [12]	.535	.375	.364	.223	.217	.161	.481	.348
LINE [61]	.549	.452	.365	.217	**.249**	.192	.499	.411
NetMF [56]	.491	.330	.377	.243	.237	.136	.456	.297
NetSMF [55]	.562	.408	.381	.228	.242	.169	.494	.391
LouvainNE [5]	.562	.395	.347	.204	.175	.114	.475	.334
ProNE [68]	.536	.443	.356	.312	.229	.200	.478	.396
NNSED [60]	.295	.243	.168	.116	.064	.035	.335	.285
MNMF [62]	.542	.362	.324	.171	.188	.102	.466	.287
BigClam [66]	.091	.066	.028	.022	.024	.015	.053	.041
SymmNMF [34]	**.596**	.397	.308	.175	.207	.088	.437	.341
HM-LDM($p=1$)	.562	.502	**.400**	**.392**	.228	**.205**	**.527**	.485
HM-LDM($p=2$)	.539	**.506**	.384	.373	.217	.183	.507	**.504**

competitive baselines for the NMI metric. For the ARI metric, we observe that our framework significantly outperforms the baselines for all of the considered networks.

Comparison with the LDM: We explore the performance of HM-LDM in comparison to the classical LDM with random effects, considering normal and squared ℓ^2-norms denoted as LDM-Re and LDM-Re-$(\ell^2)^2$, accordingly. We evaluate the models, based on link prediction and clustering tasks and report the scores in Table 5 and Table 6. The results show that despite constraining the latent space into the $D-$simplex with volumes ensuring identifiable solutions, we only observe a slight decrease in AUC-ROC scores. In contrast, the HM-LDM yields favorable NMI scores for community detection and considerably higher ARI scores when compared to classic LDM. For sufficiently large δ values (i.e. $\delta^2 = 10^3$), link-prediction performance for HM-LDM reached the one of the unconstrained LDM, but the clustering scores of the latter decrease significantly. This is since for large simplex volumes, the HM-LDM closely approximates the LDM at the expense of model and structure identifiability.

Extension to bipartite networks: We can trivially extend the HM-LDM model to account for unsigned bipartite networks [46]. Such an extension is achieved by defining and introducing a different set of latent variables for the two disjoint sets of nodes, as present in a bipartite structure. In addition, the HM-LDM(p=2) model simply extends the symmetric NMF operation, obtained for the undirected networks, to the non-symmetric NMF specification. In Fig 5, we provide the reordered adjacency matrix with respect to the community allocations defined by the learned embeddings of HM-LDM for a *Drug-Gene* [39] network ($|\mathcal{V}| = 7,341|$, $|\mathcal{E}| = 15,138$) where we observe a clear block structure. Importantly, the HM-LDM offers identifiable joint embedding representations, mixed memberships, and community discovery for bipartite networks, tasks in general considered to be non-trivial and arduous.

Table 5: HM-LDM and LDM-Re comparison for the link prediction task.

	AstroPh			*GrQc*			*Facebook*			*HepTh*		
Dimension (D)	8	16	32	8	16	32	8	16	32	8	16	32
LDM-Re	.973	.974	.979	.949	.952	.954	.993	.994	.992	.920	.923	.923
HM-LDM($p=1, \delta^2=$ identifiable)	.956	.952	.952	.944	.948	.951	.982	.979	.974	.916	.921	.924
HM-LDM($p=1, \delta^2=10^3$)	.967	.967	.965	.956	.955	.951	.985	.986	.987	.932	.931	.926
LDM-Re-$(\ell^2)^2$	.979	.978	.976	.944	.944	.945	.990	.990	.991	.913	.912	.909
HM-LDM($p=2, \delta^2=$ identifiable)	.972	.973	.963	.940	.942	.946	.992	.993	.993	.908	.910	.911
HM-LDM($p=2, \delta^2=10^3$)	.984	.983	.980	.948	.946	.946	.991	.991	.992	.920	.918	.913

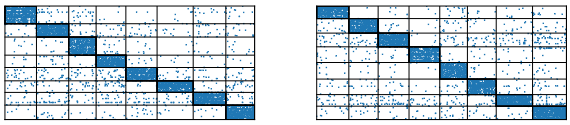
Table 6: HM-LDM and LDM-Re comparison for the clustering task.

	Amherst		*Rochester*		*Mich*		*Hamilton*	
Metric	NMI	ARI	NMI	ARI	NMI	ARI	NMI	ARI
LDM-Re	.548	.366	.391	.212	.230	.132	.491	.320
HM-LDM($p=1, \delta^2=$ identifiable)	.562	.502	.400	.392	.228	.205	.527	.485
HM-LDM($p=1, \delta^2=10^3$)	.439	.386	.308	.303	.176	.133	.405	.377
LDM-Re-$(\ell^2)^2$	.546	.370	.393	.211	.231	.137	.497	.327
HM-LDM($p=2, \delta^2=$ identifiable)	.539	.506	.384	.373	.217	.183	.507	.504
HM-LDM($p=2, \delta^2=10^3$)	.240	.133	.206	.119	.116	.056	.232	.209

3.2. *Signed Networks Experiments*

For the signed network experiments, we introduce four networks of varying sizes and structures. (**i**) *Reddit* which uses hyperlinks to create directed edges between communities belonging to the social network platform [35]. (**ii**) *wikiElec* and (**iii**) its more recent version *wikiRfa* which follow election procedures carried out through multiple timelines and convey voting information as links about users to describe positive, neutral, and opposing views for potential users to be elected administrators on Wikipedia. [37, 63]. (**iv**) *Twitter* is an undirected network with positive and negative links obtained from user tweets about the referendum concerning the reform of the Italian Constitution back in 2016 [52].

The performance of sHM-LDM is compared against seven graph representation



(a) $p = 1$, $\delta = 1$ (b) $p = 2$, $\delta = 1$

Fig. 5: *Drug-Gene* ordered adjacency matrices based on HM-LDM with $D = 8$, empirically showcasing community extraction and identification extended to bipartite networks.

Table 7: Binary operators considered for designing feature vectors (edge features). The notation, $f(v)_d$ denotes the d'th coordinate of the embedding vector of node v.

Operator	**Symbol**	**Definition**
Average	$\boxplus$	$[f(u) \boxplus f(v)]_d = (f(u)_d + f(v)_d)/2$
Hadamard	$\boxdot$	$[f(u) \boxdot f(v)]_d = f(u)_d \cdot f(v)_d$
Weighted-L1	$\|\cdot\|_1$	$\|f(u) - f(v)\|_{1_d} = \|f(u)_d - f(v)_d\|$
Weighted-L2	$\|\cdot\|_2$	$\|f(u) - f(v)\|_{2_d} = \|f(u)_d - f(v)_d\|^2$
Concatenate	$\oplus$	$[f(u) \oplus f(v)]_d = \big(f(u)_d,\ f(v)_d\big)$

Table 8: Area Under Curve (AUC-ROC) scores for representation size of $D = 8$ and δ values ensuring identifiability. ("x" denotes a baseline that was unable to run due to high memory/runtime complexity)

	WikiElec			*WikiRfa*			*Twitter*			*Reddit*		
Task	$p@n$	$p@z$	$n@z$	$p@n$	$p@z$	$n@z$	$p@n$	$p@z$	$n@z$	$p@n$	$p@z$	$n@z$
POLE [24]	.809	.896	.853	.904	.921	.767	.965	.902	.922	x	x	x
SLF [65]	**.888**	.954	**.952**	**.971**	.963	.961	.914	.877	.968	**.729**	**.955**	.968
SiGAT [21]	.874	.775	.754	.944	.766	.792	**.998**	.875	.963	.707	.682	.712
SIDE [30]	.728	.866	.895	.869	.861	.908	.799	.843	.910	.653	.830	.892
SignNet [26]	.841	.774	.635	.920	.736	.717	.968	.719	.891	.646	.547	.623
SLDM [48]	.876	**.969**	.936	.963	**.982**	.963	.986	.962	.973	.648	.951	.975
SLIM [48]	.862	.965	.935	.956	.980	.960	.988	**.963**	.972	.667	**.955**	**.978**
sHM-LDM(p=1)	.872	.963	.938	.959	.977	.963	.978	.958	**.976**	.642	.951	.977
sHM-LDM(p=2)	.872	.966	.937	.960	.975	**.964**	.977	.958	.973	.610	.953	.976

learning baselines, eligible for analyzing signed networks: (**i**) POLE [24] where embeddings are based on the decomposition of an auto-covariance matrix created through signed random walks, (**ii**) SLF [65] that creates representations based on latent factors capable of describing both positive and negative connections, (**iii**) SiGAT [21] a graph neural network model learning node embeddings through a graph attention mechanism, (**iv**) SIDE [30] utilizing truncated random walks under a general likelihood expression for signed relationships modeling both positive and negative ties, (**v**) SignNet [26] a deep neural network using a similarity measure through the Hadamard product able to describe signed proximity between a pair of nodes, (**vi**) *SLDM* and (**vii**) *SLIM* models [48] which define an unconstrained and a constrained to polytopes latent distance model, respectively. Both of these two models utilize the Skellam distribution as the sHM-LDM which constrains the model to the $D-$simplex while *SLIM* operates on the inferred convex-hull of the latent space.

3.3. *Signed Link prediction*

We follow the same evaluation procedure as in [48] and define two settings considering link prediction, in order to benchmark sHM-LDM's predictive capability

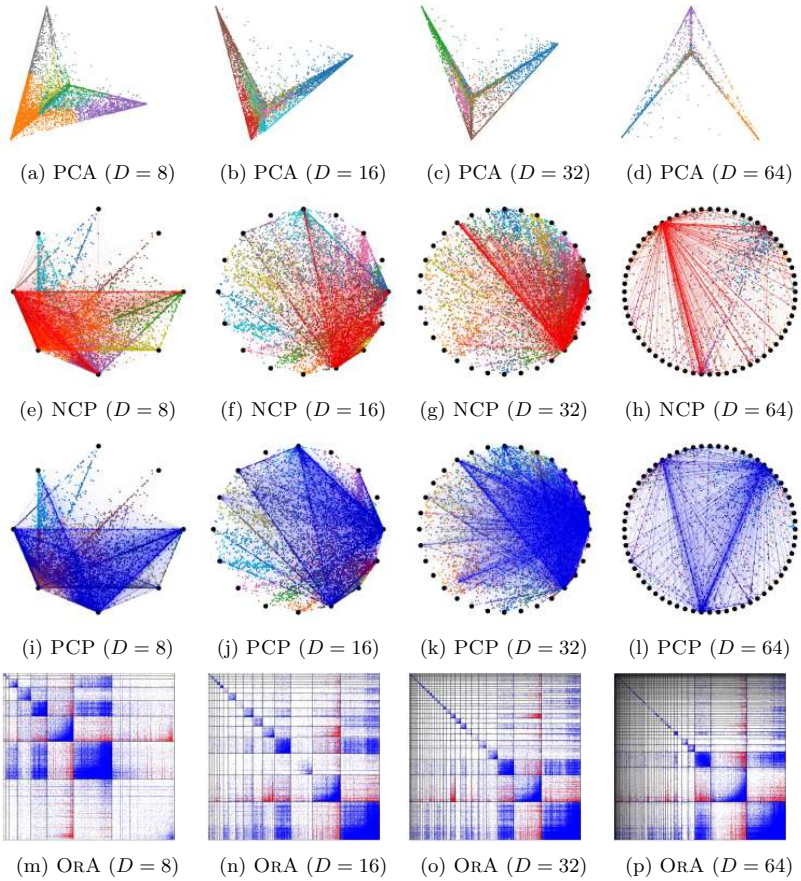
against the considered baselines. For that, we randomly choose 20% of the total network links/cases (both positive and negative) which are then zeroed out with the constraint that the residual signed network stays connected. Furthermore, an equal size of disconnected pairs in the original networks is also drawn to act as the controls in the prediction tasks. The combined samples of removed links and drawn controls define the test set for each network. All models are trained on the residual networks for each dataset.

Performance evaluation: For our methods, as well as, for *SLDM* and *SLIM* we learn a logistic regression model with inputs given by both the rates and log-rates, as defined by the Skellam distribution, i.e. $\chi_{ij} = \left[\lambda_{ij}^{+}, \lambda_{ij}^{-}, \log \lambda_{ij}^{+}, \log \lambda_{ij}^{-}\right]$. It is argued in [48] that the Skellam distribution operates on both the ratios and products of the rates during inference. Consequently, training a logistic regression based on rates and log rates allow for learning linear and non-linear mappings based on both the rates as well as their products and ratios due to the log transformation. For the performance evaluation of the baselines, we consider five binary operators, as established in the GRL literature. These include the {average, Hadamard product, weighted L1, weighted L2, concatenate}, as shown in Table 7. These are utilized to construct five different feature vectors, used to train multiple logistic regression models for each task. For every baseline defining multiple feature vectors, we choose the logistic regression model that returns the maximum performance for each individual task. Lastly, for each link prediction task, we consider the robust against class imbalance metric, area-under-curve of the receiver operating characteristic (AUC-ROC).

Task 1: Link sign prediction. For the first task we consider only the links/cases of the test set for each network. After training, each model is provided with the test set link pairs and evaluated in its ability to predict the sign of the removed links. The AUC-ROC results are summarized in Table 8 where the link sign prediction is represented as $p@n$. We mostly observe favorable or on-par results and performance against the baselines. More specifically, comparing to the *SLDM* and *SLIM*, our models despite defining a more constrained latent space (recall that $\mathbf{A} = \delta \cdot \mathbf{I}$ for sHM-LDM) the obtained results shows identical or on-par performance.

Task 2: Signed link prediction. The second task is more difficult and evaluates the performance of a model in its ability to both predict the sign, as well as, the presence of a link. For that, the whole test set is used to create two test subsets. The first contains the controls and positive links while the second the controls and the negative links. The models then are asked to distinguish controls from positive cases and controls from negative cases, respectively. We denote these tasks accordingly as $p@z$ and $n@z$ and AUC-ROC scores are provided in Table 8. Once more, the *sHM-LDM* frameworks provide favorable or on-par performance against the baselines and especially to the *SLDM* and *SLIM* models.

Visualizations: The inferred simplex of sHM-LDM extracts information about



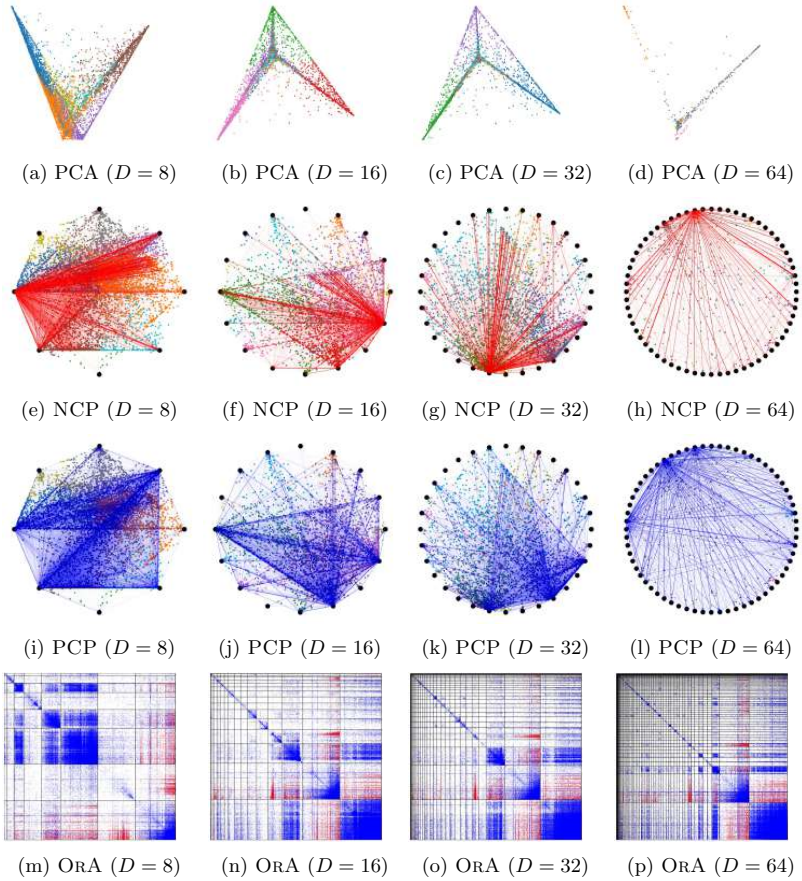
(a) PCA ($D = 8$) (b) PCA ($D = 16$) (c) PCA ($D = 32$) (d) PCA ($D = 64$)

(e) NCP ($D = 8$) (f) NCP ($D = 16$) (g) NCP ($D = 32$) (h) NCP ($D = 64$)

(i) PCP ($D = 8$) (j) PCP ($D = 16$) (k) PCP ($D = 32$) (l) PCP ($D = 64$)

(m) ORA ($D = 8$) (n) ORA ($D = 16$) (o) ORA ($D = 32$) (p) ORA ($D = 64$)

Fig. 6: **sHM-LDM(p=2)**: *Twitter* Network—Inferred simplex visualizations and ordered adjacency matrices for various dimensions D and with simplex side lengths δ ensuring identifiability. The first row shows the latent space projection to the first two Principal Components—The second row provides a Negative Circular Plot (NCP) with red lines showcasing negative links between nodes—The third row shows a Positive Circular Plot (PCP) with the blue lines denoting positive links between node pairs—The fourth and final row shows the Ordered Adjacency (ORA) matrices sorted based on the memberships $\mathbf{w}_i$, in terms of maximum simplex corner responsibility, and internally according to the magnitude of the corresponding corner assignment for their reconstruction.



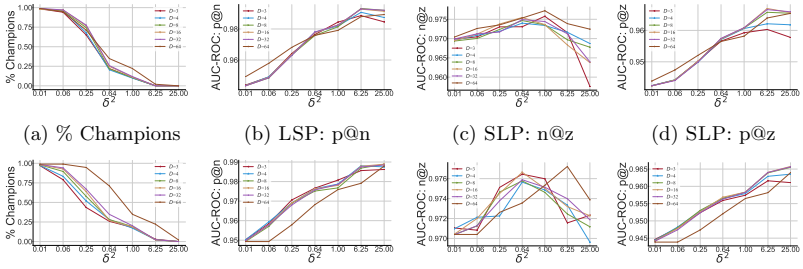
(a) PCA ($D = 8$) (b) PCA ($D = 16$) (c) PCA ($D = 32$) (d) PCA ($D = 64$)

(e) NCP ($D = 8$) (f) NCP ($D = 16$) (g) NCP ($D = 32$) (h) NCP ($D = 64$)

(i) PCP ($D = 8$) (j) PCP ($D = 16$) (k) PCP ($D = 32$) (l) PCP ($D = 64$)

(m) ORA ($D = 8$) (n) ORA ($D = 16$) (o) ORA ($D = 32$) (p) ORA ($D = 64$)

Fig. 7: **sHM-LDM(p=1)**: *Twitter* Network—Inferred simplex visualizations and ordered adjacency matrices for various dimensions D and with simplex side lengths δ ensuring identifiability. The first row shows the latent space projection to the first two Principal Components—The second row provides a Negative Circular Plot (NCP) with red lines showcasing negative links between nodes—The third row shows a Positive Circular Plot (PCP) with the blue lines denoting positive links between node pairs—The fourth and final row shows the Ordered Adjacency (ORA) matrices sorted based on the memberships $\mathbf{w}_i$, in terms of maximum simplex corner responsibility, and internally according to the magnitude of the corresponding corner assignment for their reconstruction.

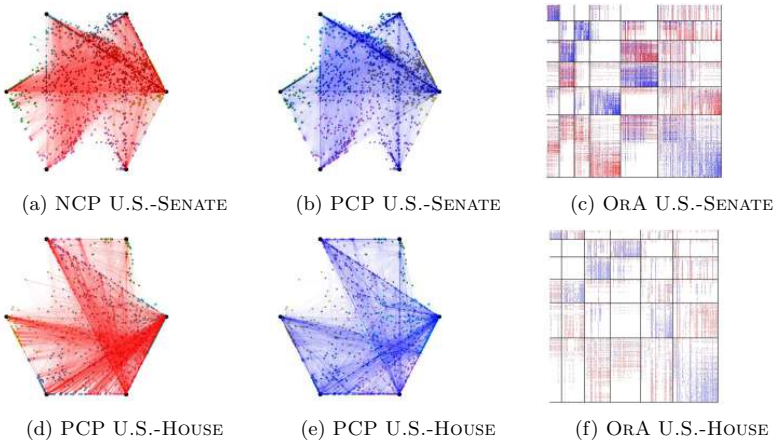


(a) % Champions (b) LSP: p@n (c) SLP: n@z (d) SLP: p@z

(e) % Champions (f) LSP: p@n (g) SLP: n@z (h) SLP: p@z

Fig. 8: **sHM-LDM**: *Twitter* Network—Performance characteristics across different dimensions D in terms of various values δ^2 (simplex size). The first column shows the total community champions (%) across dimensions for sHM-LDM—The second column provides the Link Sign Prediction (LSP) performance for the task of inferring the sign of the test set links (p@n)—The third and fourth columns describe the performance for the Signed Link Prediction (SLP) tasks, distinguishing between negatively related and disconnected nodes (n@z), as well as, positively connected to disconnected nodes (p@z), respectively. Top row: $p = 2$ model specification. Bottom row $p = 1$ model specification.

node memberships to distinct aspects of the network. Similar to [48], we provide visualizations regarding the latent space as projected to the first two principal components and include circular plots describing the simplex and node memberships in two dimensions. Specifically, each corner of the simplex is positioned to the border of a circle, every $\text{rad}_k = \frac{2\pi}{D}$ radians, with D being the number of the simplex corners. Furthermore, we provide the re-ordered adjacency matrices based on the inferred memberships for various dimensions. Visualizations for the *Twitter* are provided in Fig. 6 and Fig. 7 for sHM-LDM($p = 2$) and sHM-LDM($p = 1$) models, respectively. For both models, visualizations are available for different dimensions while we see how the model successfully uncovers distinct aspects of the network when the simplex side length δ ensures identifiability. From the circular plots enriched with the corresponding negative (red lines) and positive (blue lines) links, we observe that the models always uncover simplex corners to act as dislike (high negative in-degree) and like (high positive in-degree) profiles of the network. We also observe controversial network profiles, sharing a high degree of both negative and positive connections. For the ordered adjacency matrices of the two models, we can observe successful structure extraction and discovery, and as we increase the dimensionality of the simplex structure it becomes finer and finer. Lastly, we also obtain simplex corners for the inferred simplex containing not-so-intensely connected nodes. This comes as a validation of stochastic equivalence presence that the



(a) NCP U.S.-Senate (b) PCP U.S.-Senate (c) OrA U.S.-Senate

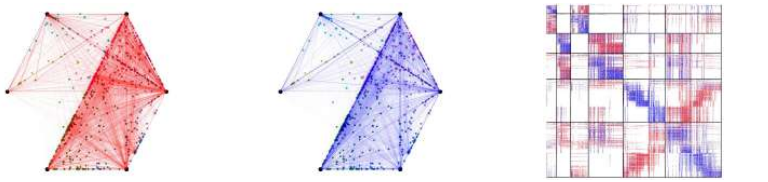
(d) PCP U.S.-House (e) PCP U.S.-House (f) OrA U.S.-House

Fig. 9: **sHM-LDM**(p=2): Inferred simplex visualizations and ordered adjacency matrices for a $D = 6$ dimensional simplex with side lengths δ ensuring identifiability. The first column provides a Negative Circular Plot (NCP) with red lines showcasing negative links between nodes—The second column shows a Positive Circular Plot (PCP) with the blue lines denoting positive links between node pairs—The third and final column shows the Ordered Adjacency (OrA) matrices ordered based on the memberships, in terms of maximum simplex corner responsibility, and internally according to the magnitude of the corresponding corner assignment for their reconstruction. Top row: U.S.-House. Bottom row U.S.-Senate.

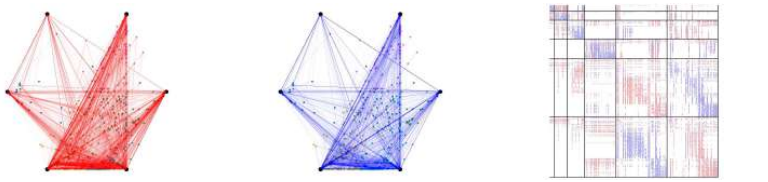
sHM-LDM framework can express.

Simplex size and performance evaluation: In Fig. 8 we provide performance characteristics against various dimensions D as a function of δ^2 for sHM-LDM(p=2) and sHM-LDM(p=1) models, respectively. The first column shows the percentage of champion nodes as defined by the model whereas expected smaller simplex volumes lead to a higher percentage of hard-clustered nodes. In addition, it is clear that the dimensionality in sHM-LDM(p=1) has a bigger effect on the node champions than for the sHM-LDM(p=2) case. The last three columns showcase the performance across the p@z, n@z, and p@z tasks respectively. Comparing to the results of theHM-LDM we observe for the signed networks and sHM-LDM that the performance is not affected to the same degree by the shrinkage of the latent space (the maximum case is present in the p@n task accounting to just a 4% decrease).

Extension to signed bipartite networks: Here, similar to the unsigned network study, we extend the analysis to bipartite signed networks for sHM-LDM. The extension is again trivial by defining two sets of latent variables describing the



(a) NCP U.S.-Senate (b) PCP U.S.-Senate (c) OrA U.S.-Senate



(d) PCP U.S.-House (e) PCP U.S.-House (f) OrA U.S.-House

Fig. 10: **sHM-LDM**(p=1): Inferred simplex visualizations and ordered adjacency matrices for a $D = 6$ dimensional simplex with side lengths δ ensuring identifiability. The first column provides a Negative Circular Plot (NCP) with red lines showcasing negative links between nodes—The second column shows a Positive Circular Plot (PCP) with the blue lines denoting positive links between node pairs—The third and final column shows the Ordered Adjacency (OrA) matrices ordered based on the memberships, in terms of maximum simplex corner responsibility, and internally according to the magnitude of the corresponding corner assignment for their reconstruction. Top row: U.S.-House. Bottom row U.S.-Senate.

two disjoint groups of nodes, as present in bipartite structures. In addition, we introduce four sets of random effects defining again node social and antisocial reach but now respecting target and source roles of the nodes in the corresponding networks links. We introduce two signed bipartite networks, *U.S.-House* [11] ($|\mathcal{V}| = 1796$, $|\mathcal{E}^+| = 61678$, $|\mathcal{E}^-| = 52619$, Density=0.1734), and *U.S.-Senate* [11] ($|\mathcal{V}| = 1201$, $|\mathcal{E}^+| = 14964$, $|\mathcal{E}^-| = 12096$, Density=0.1769), regarding voting records for proposed bills as made by the U.S. House of Representatives and the U.S. Senate, accordingly. For these networks, the first (rows) of the disjoint sets of nodes refer to the bills while the second (columns) to representatives or senators, accordingly. In Figs 9 and 10, we provide the Positive Circular Plots PCP, Negative Circular Plots NCP, and Ordered Adjacency Matrices OrA for the corresponding networks and for both sHM-LDM(p=2) and sHM-LDM(p=1) frameworks, respectively. We witness how the sHM-LDM framework generalizes to the study of bipartite networks, successfully uncovering distinct network aspects and profiles, that convey information about both homophily, as well as, animosity being present in the network.

4. Complexity analysis

The proposed HM-LDM and its signed extension sHM-LDM belong to the family of latent distance models and thus require the calculation of the all-pairs distance matrix. This scales as $\mathcal{O}(N^2)$ in time and memory, making large-scale network analysis infeasible. To alleviate that problem we consider unbiased estimations of the log-likelihood through a random sampling approach. More specifically, in every model iteration, a set of network nodes, $S \subseteq \mathcal{V}$, is sampled (with replacement) and gradient steps are taken based on the log-likelihood of the block defined by the sampled node set. This effectively reduces the complexity of the models to $\mathcal{O}(S^2)$ both in time and memory. Another option is the case-control approach [57] scaling by the number of network edges as $\mathcal{O}(E)$. Lastly, the Hierarchical Block Distance Model (HBDM) [49] is an attractive alternative option where gradient steps over the model parameters are based on a hierarchical approximation of the likelihood of the whole network. The HBDM model scales linearithmicly as $\mathcal{O}(N \log N)$ both in space and time while also offering hierarchical characterizations of structures at multiple scales.

5. Conclusion and future work

In this study, we have presented the HM-LDM reconciling graph representation learning and latent community detection. We extended the model to account for signed networks and showed that a minimum volume approach could uncover distinct profiles in social networks while ensuring model identifiability. Both presented frameworks were formulated to include a Euclidean as well as a squared Euclidean norm. For the latter, a direct relationship to an Eigenmodel in both the case of unsigned and signed networks was shown. Furthermore, by controlling the volume of the simplex by the magnitude of δ, a sufficiently reduced simplex leads to unique representations. For unsigned networks, this resulted in the hard clustering of nodes to communities when the simplex was sufficiently contracted. Notably, the generalization to signed networks facilitated the extraction of distinct network profiles representing positive interactions and animosity. In regimes where HM-LDM and sHM-LDM provide unique representations, we observed favorable link prediction performance and the ability to order the adjacency matrix based on prominent latent communities and distinct profiles. Notably, the proposed HM-LDM combines network homophily and transitivity properties with latent community detection enabling explicit control of soft and hard assignment through the volume of the induced simplex. Importantly, the extended sHM-LDM merges homophily and heterophily properties to account for positive and negative ties as present in signed networks. To further evaluate the performance of HM-LDM and sHM-LDM, future work should compare them against classical non-embedding methods such as the Degree Corrected Stochastic Block Model (DC-SBM) [29] or the Mixed Membership Stochastic Block Model (MM-SBM) [1], as well as, a Stochastic Block Model accounting for signed networks [28].

Acknowledgements

We would like to express sincere appreciation and thank the reviewers for their constructive feedback and their insightful comments. We gratefully acknowledge the Independent Research Fund Denmark for supporting this work [grant number: 0136-00315B].

References

[1] Airoldi, E. M., Blei, D. M., Fienberg, S. E., and Xing, E. P., Mixed membership stochastic blockmodels, *J Mach Learn Res* **9** (2008) 1981–2014.

[2] Atay, F. and Tunçel Gölpek, H., On the spectrum of the normalized laplacian for signed graphs: Interlacing, contraction, and replication, *Linear Algebra and its Applications* **442** (2014) 165–177.

[3] Ball, B., Karrer, B., and Newman, M. E. J., An efficient and principled method for detecting communities in networks, *CoRR* **abs/1104.3590** (2011).

[4] Beentjes, S. V. and Khamseh, A., Higher-order interactions in statistical physics and machine learning: A model-independent solution to the inverse problem at equilibrium, *Phys. Rev. E* **102** (2020) 053314.

[5] Bhowmick, A. K., Meneni, K., Danisch, M., Guillaume, J.-L., and Mitra, B., LouvainNE: Hierarchical louvain method for high quality and scalable network embedding, in *WSDM* (2020), pp. 43–51.

[6] Bioucas-Dias, J. M., Plaza, A., Dobigeon, N., Parente, M., Du, Q., Gader, P., and Chanussot, J., Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **5** (2012) 354–379.

[7] Büeler, B., Enge, A., and Fukuda, K., Exact volume computation for polytopes: a practical study, in *Polytopes—combinatorics and computation* (Springer, 2000), pp. 131–154.

[8] Çelikkanat, A. and Malliaros, F. D., Exponential family graph embeddings, in *AAAI* (2020), pp. 3357–3364.

[9] Chakraborty, T., Dalmia, A., Mukherjee, A., and Ganguly, N., Metrics for community analysis: A survey (2016).

[10] Cutler, A. and Breiman, L., Archetypal analysis, *Technometrics* **36** (1994) 338–347.

[11] Derr, T., Johnson, C., Chang, Y., and Tang, J., Balance in signed bipartite networks, in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19 (Association for Computing Machinery, New York, NY, USA, 2019), ISBN 9781450369763, p. 1221–1230, doi:10.1145/3357384.3358009, `https://doi.org/10.1145/3357384.3358009`.

[12] Grover, A. and Leskovec, J., Node2Vec: Scalable feature learning for networks, in *KDD* (2016), pp. 855–864.

[13] Hamilton, W. L., Graph representation learning, *Synthesis Lectures on Artificial Intelligence and Machine Learning* **14** 1–159.

[14] Hamilton, W. L., Ying, R., and Leskovec, J., Inductive representation learning on large graphs, in *NIPS* (2017).

[15] Hamilton, W. L., Ying, R., and Leskovec, J., Representation learning on graphs: Methods and applications, *IEEE Data Eng. Bull.* **40** (2017) 52–74.

[16] Handcock, M. S., Raftery, A. E., and Tantrum, J. M., Model-based clustering for social networks, *J R Stat Soc Ser A Stat Soc.* **170** (2007) 301–354.

[17] Hart, Y., Sheftel, H., Hausser, J., Szekely, P., Ben-Moshe, N. B., Korem, Y., Tendler,

A., Mayo, A. E., and Alon, U., Inferring biological tasks using pareto analysis of high-dimensional data, *Nature methods* **12** (2015) 233–235.

[18] Hoff, P. D., Bilinear mixed-effects models for dyadic data, *JASA* **100** (2005) 286–295.

[19] Hoff, P. D., Modeling homophily and stochastic equivalence in symmetric relational data, in *NIPS* (2007), p. 657–664.

[20] Hoff, P. D., Raftery, A. E., and Handcock, M. S., Latent space approaches to social network analysis, *JASA* **97** (2002) 1090–1098.

[21] Huang, J., Shen, H., Hou, L., and Cheng, X., Signed graph attention networks, in *ICANN 2019: Workshop and Special Sessions* (2019), pp. 566–577.

[22] Huang, J., Shen, H., Hou, L., and Cheng, X., SDGNN: Learning node representation for signed directed networks, *AAAI* **35** (2021) 196–203.

[23] Huang, K., Sidiropoulos, N. D., and Swami, A., Non-negative matrix factorization revisited: Uniqueness and algorithm for symmetric decomposition, *IEEE Trans. Signal Process* **62** (2014) 211–224.

[24] Huang, Z., Silva, A., and Singh, A., POLE: Polarized embedding for signed networks, *WSDM* (2022) 390–400.

[25] Hummon, N. P. and Doreian, P., Some dynamics of social balance processes: bringing heider back into balance theory, *Social Networks* **25** (2003) 17–49.

[26] Islam, M. R., Aditya Prakash, B., and Ramakrishnan, N., SIGNet: Scalable embeddings for signed networks, in *Advances in Knowledge Discovery and Data Mining*, eds. Phung, D., Tseng, V. S., Webb, G. I., Ho, B., Ganji, M., and Rashidi, L. (Springer International Publishing, Cham, 2018), pp. 157–169.

[27] Jianbo Shi and Malik, J., Normalized cuts and image segmentation, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22** (2000) 888–905.

[28] Jiang, J., Stochastic blockmodel and exploratory analysis in signed networks, *Physical Review E* **91** (2015).

[29] Karrer, B. and Newman, M. E., Stochastic blockmodels and community structure in networks, *Physical review E* **83** (2011) 016107.

[30] Kim, J., Park, H., Lee, J.-E., and Kang, U., SIDE: Representation learning in signed directed networks, in *Proceedings of the 2018 World Wide Web Conference* (International World Wide Web Conferences Steering Committee, 2018), p. 509–518.

[31] Kingma, D. P. and Ba, J., Adam: A method for stochastic optimization, in *ICLR* (2015).

[32] Kipf, T. N. and Welling, M., Semi-supervised classification with graph convolutional networks, in *ICLR* (2017).

[33] Krivitsky, P. N., Handcock, M. S., Raftery, A. E., and Hoff, P. D., Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models, *Social Networks* **31** (2009) 204 – 213.

[34] Kuang, D., Ding, C., and Park, H., Symmetric nonnegative matrix factorization for graph clustering, in *SDM* (2012).

[35] Kumar, S., Hamilton, W. L., Leskovec, J., and Jurafsky, D., Community interaction and conflict on the web, in *WWW* (2018), pp. 933–943.

[36] Lee, D. D. and Seung, H. S., Learning the parts of objects by nonnegative matrix factorization, *Nature* **401** (1999) 788–791.

[37] Leskovec, J., Huttenlocher, D., and Kleinberg, J., Predicting positive and negative links in online social networks, in *WWW* (2010), p. 641–650.

[38] Leskovec, J., Kleinberg, J., and Faloutsos, C., Graph evolution: Densification and shrinking diameters, *TKDD* **1** (2007).

[39] Leskovec, J. and Krevl, A., SNAP Datasets: Stanford large network dataset collection (2014).

[40] Leskovec, J. and Mcauley, J. J., Learning to discover social circles in ego networks, *NIPS* (2012) 539–547.
[41] Mao, X., Sarkar, P., and Chakrabarti, D., On mixed memberships and symmetric nonnegative matrix factorizations, in *ICML*, Vol. 70 (2017).
[42] Miao, L. and Qi, H., Endmember extraction from highly mixed data using minimum volume constrained nonnegative matrix factorization, *IEEE Transactions on Geoscience and Remote Sensing* **45** (2007) 765–777.
[43] Mucha, P. and Porter, M., Social structure of facebook networks, *Physica A: Statistical Mechanics and its Applications* **391** (2012) 4165–4180.
[44] Muolo, R., Gallo, L., Latora, V., Frasca, M., and Carletti, T., Turing patterns in systems with high-order interactions, *Chaos, Solitons & Fractals* **166** (2023) 112912.
[45] Mørup, M. and Kai Hansen, L., Archetypal analysis for machine learning, in *Workshop on Machine Learning for Signal Processing* (2010), pp. 172–177.
[46] Nakis, N., Çelikkanat, A., and Mørup, M., Scalable hierarchical embeddings of complex networks (2022), `https://openreview.net/pdf?id=U-GB_gONqbo`.
[47] Nakis, N., Çelikkanat, A., and Mørup, M., Hm-ldm: A hybrid-membership latent distance model, in *Complex Networks and Their Applications XI*, eds. Cherifi, H., Mantegna, R. N., Rocha, L. M., Cherifi, C., and Miccichè, S. (Springer International Publishing, Cham, 2023), ISBN 978-3-031-21127-0, pp. 350–363.
[48] Nakis, N., Çelikkanat, A., Boucherie, L., Djurhuus, C., Burmester, F., Holmelund, D., Frolcová, M., and Mørup, M., Characterizing polarization in social networks using the signed relational latent distance model, in *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics* (2023).
[49] Nakis, N., Çelikkanat, A., Jørgensen, S. L., and Mørup, M., A hierarchical block distance model for ultra low-dimensional graph representations (2022).
[50] Newman, M. E. J., The structure and function of complex networks, *SIAM Review* **45** (2003) 167–256.
[51] Ng, A. Y., Jordan, M. I., and Weiss, Y., On spectral clustering: Analysis and an algorithm, in *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, NIPS'01 (MIT Press, Cambridge, MA, USA, 2001), p. 849–856.
[52] Ordozgoiti, B., Matakos, A., and Gionis, A., Finding large balanced subgraphs in signed networks, in *Proceedings of The Web Conference 2020* (2020), p. 1378–1388.
[53] Ou, M., Cui, P., Pei, J., Zhang, Z., and Zhu, W., Asymmetric transitivity preserving graph embedding, in *KDD* (2016), pp. 1105–1114.
[54] Perozzi, B., Al-Rfou, R., and Skiena, S., Deepwalk: Online learning of social representations, in *KDD* (2014), p. 701–710.
[55] Qiu, J., Dong, Y., Ma, H., Li, J., Wang, C., Wang, K., and Tang, J., NetSMF: Large-scale network embedding as sparse matrix factorization, in *WWW* (2019).
[56] Qiu, J., Dong, Y., Ma, H., Li, J., Wang, K., and Tang, J., Network embedding as matrix factorization: Unifying DeepWalk, LINE, PTE, and Node2Vec, in *WSDM* (2018), pp. 459–467.
[57] Raftery, A. E., Niu, X., Hoff, P. D., and Yeung, K. Y., Fast inference for the latent space network model using a case-control approximate likelihood, *Journal of Computational and Graphical Statistics* **21** (2012) 901–919.
[58] Ryan, C., Wyse, J., and Friel, N., Bayesian model selection for the latent position cluster model for social networks, *Network Science* **5** (2017) 70–91.
[59] Skellam, J. G., The frequency distribution of the difference between two poisson variates belonging to different populations., *Journal of the Royal Statistical Society. Series A (General)* **109** (1946) 296–296.

[60] Sun, B.-J., Shen, H., Gao, J., Ouyang, W., and Cheng, X., A non-negative symmetric encoder-decoder approach for community detection, in *CIKM* (2017).
[61] Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., and Mei, Q., LINE: Large-scale information network embedding, in *WWW* (2015), pp. 1067–1077.
[62] Wang, X., Cui, P., Wang, J., Pei, J., Zhu, W., and Yang, S., Community preserving network embedding, in *AAAI* (2017).
[63] West, R., Paskov, H. S., Leskovec, J., and Potts, C., Exploiting social network structure for person-to-person sentiment analysis, *TACL* **2** (2014) 297–310.
[64] Wind, D. K. and Mørup, M., Link prediction in weighted networks, in *Workshop on Machine Learning for Signal Processing* (2012), pp. 1–6.
[65] Xu, P., Wu, J., Hu, W., and Du, B., Link prediction with signed latent factors in signed social networks, *Proceedings of the Acm Sigkdd International Conference on Knowledge Discovery and Data Mining* (2019) 1046–1054.
[66] Yang, J. and Leskovec, J., Overlapping community detection at scale: A nonnegative matrix factorization approach, in *WSDM* (2013).
[67] Zhang, D., Yin, J., Zhu, X., and Zhang, C., Network representation learning: A survey, *IEEE Trans. Big Data* **6** (2020).
[68] Zhang, J., Dong, Y., Wang, Y., Tang, J., and Ding, M., Prone: Fast and scalable network representation learning, in *IJCAI* (2019).
[69] Zhuang, L., Lin, C.-H., Figueiredo, M. A., and Bioucas-Dias, J. M., Regularization parameter selection in minimum volume hyperspectral unmixing, *IEEE Transactions on Geoscience and Remote Sensing* **57** (2019) 9858–9877.
[70] Çelikkanat, A., Nakis, N., and Mørup, M., Piecewise-velocity model for learning continuous-time dynamic node representations (2022).

Time to Cite: Modeling Citation Networks using the Dynamic Impact Single-Event Embedding Model

Nikolaos Nakis
Technical University of Denmark
Kongens Lyngby 2800, Denmark
nnak@dtu.dk

Abdulkadir Çelikkanat
Technical University of Denmark
Kongens Lyngby 2800, Denmark
abce@dtu.dk

Louis Boucherie
Technical University of Denmark
Kongens Lyngby 2800, Denmark
louibo.dk

Sune Lehmann
Technical University of Denmark
Kongens Lyngby 2800, Denmark
sljo@dtu.dk

Morten Mørup
Technical University of Denmark
Kongens Lyngby 2800, Denmark
mmor@dtu.dk

Abstract

Understanding the structure and dynamics of scientific research, i.e., the science of science (SciSci), has become an important area of research in order to address imminent questions including how scholars interact to advance science, how disciplines are related and evolve, and how research impact can be quantified and predicted. Central to the study of SciSci has been the analysis of citation networks. Here, two prominent modeling methodologies have been employed: one is to assess the citation impact dynamics of papers using parametric distributions, and the other is to embed the citation networks in a latent space optimal for characterizing the static relations between papers in terms of their citations. Interestingly, citation networks are a prominent example of what we denote as single-event dynamic networks, i.e., networks for which each dyad only has a single event (i.e., the point in time of citation). We presently propose a novel likelihood function for the characterization of such single-event networks. Using this likelihood, we further propose the Dynamic Impact Single-Event Embedding model (DISEE). The DISEE model characterizes the scientific interactions in terms of a latent distance model in which forces (strength of the interaction) can be reparameterized to be proportional to the product of the masses of the interacting entities. To account for the time-varying impact, the mass of a contribution being used is time-dependent based on flexible parametric representations of scientific impact. We highlight the proposed approach on several real networks of scientific collaboration finding that the DISEE well reconciles static latent distance network embedding approaches with classical dynamic impact assessments of citation networks.

1 Introduction

Networks are widespread data structures and represent the most natural means of expressing complex systems. They appear across various scientific domains, encompassing fields such as physics, sociology, science of science. biology, and more. Within these disciplines, networks are used

Preprint. Under review.

to describe a multitude of interactions and systems, such as spin glasses in physics, friendship interactions in sociology, scholarly collaborations in academia, protein-to-protein interactions in biology, and structural and functional brain connectivity in neuroscience, among many others [62]. Given their complexity and high-dimensional discrete nature, accurately characterizing the structure of networks is regarded as a non-trivial and challenging task with a plethora of methodologies and tools being developed to examine these networks and seek to gain insights into their underlying structures. These tools are used for several downstream tasks, including link/relation prediction [56], node classification and clustering [27, 31], and community detection [22].

The abundance of scientific data has established the science of science (SciSci) as a vital tool in understanding scientific research, as well as, in predicting future outcomes, research directions, and the overall evolution of science [23]. More specifically, SciSci studies the methods of science itself, searching for answers to important questions such as how scholars interact to advance science, how different disciplinary boundaries are removed, and how research impact can be quantified and predicted. SciSci is an interdisciplinary field with various prominent research directions including but not limited to, scientific novelty and innovation quantification [93, 89, 100, 12, 24], analysis of career success dynamics of scholars [71, 25, 50, 17, 68, 53, 54], characterization of scientific collaborations [13, 49, 102, 58, 7] as well as citation and research impact dynamics [95, 38, 26, 74, 97, 29, 96].

A major focus has been given to the understanding of SciSci through the lens of complex network analysis, studying the structural properties and dynamics, of the naturally occurring graph data describing SciSci. These include collaboration networks describing how scholars cooperate to advance various scientific fields. In particular, pioneering works [65, 63, 64] have analyzed various network statistics such as degree distribution, clustering coefficient, and average shortest paths. Furthermore, citation networks define an additional prominent case where graph structure data describe SciSci. Citation networks, essentially describe the directed relationships of papers (nodes) with an edge occurring between a dyad if paper A cites paper B, e.g. A$\rightarrow$B. Studies focusing on citation networks have shown power-law and exponential family degree distributions [76], sub-field community structures [32], and tree-like backbone topologies [33]. Lastly, bipartite network structures can emerge by defining networks describing author-paper relationships, including indirect author connections through a collaboration paper or through their citing patterns [28, 107, 3].

In this paper, we focus on citation networks that allow for paper impact characterization. Notably, such networks are directed and dynamic ideally having an upper triangular adjacency matrix when nodes are sorted by time due to the time-causal structure of citations (i.e., new papers can only cite past papers).

Initial works for paper impact quantification utilized classical machine learning methods on various scholarly features, as well as paper textual information. Methods used to estimate future citations included linear/logistic regression, k-nearest neighbors, support vector machines, random forests, and many more [14, 87, 4, 85, 106, 42]. These studies focused primarily on carefully designing and including proper features to be used for the impact prediction task. Prominent examples are various author and venue-based metrics such as the H-index, impact factor (IF), and, some network-based characteristics such as the centrality of authors and periodicals. In addition, improvements in the predicting scores were achieved by introducing clusters based on citation patterns of papers in their initial stage [14], as well as, modeling the interdisciplinarity of venues and authors which was achieved via the Jensen-Shannon divergence [4]. While these methods attracted lots of attention, they have a major limitation where papers with very similar features define much different citation distributions and attention patterns that are not characterized.

Later works tried to define impact on the paper level by treating the accumulation of citations through time as a time series. In the original work of [77], Redner proposed a log-normal distribution to fit the cumulative citation distribution for papers published during a 110 year period in Physical Review. This was followed by [20] using a shifted power-law distribution on the same networks. Furthermore, another widely used distribution modeling citation dynamics is the Tsallis distribution proposed in [94]. The log-normal and Tsallis distribution share a lot of similarities but in literature, the log-normal is preferred due to its simplicity. Later works combined the important characteristic of preferential attachment with the log-normal distribution [103, 98, 99], as well as, the Poisson process [84, 104].

Many studies have focused on creating detailed maps of science based on citation networks in order to unfold and visualize underlying structures revealing communities and cross-disciplinary interactions [8, 79, 6, 9, 88, 69]. These prior works concentrated their attention on citation and interactions

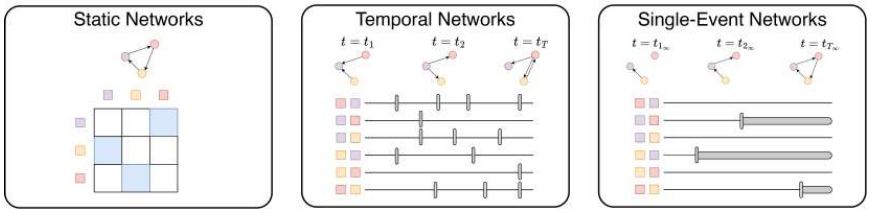


Figure 1: Examples of three different types of networks based on their temporal structure. Round points represent network nodes, square points make up the corresponding colored node dyads, arrows represent directed relationships between two nodes, vertical lines represent events, and black lines are the timelines while grey bold lines show that a link (event) appeared once and cannot be observed again. *Left panel:* Static networks where links occur once and there is no temporal information available. *Middle panel:* Temporal networks where links are events in time and can be observed multiple times along the timeline. *Right panel:* Single-event networks (SENs) where links appear in a temporal manner but they can occur only once for each dyad, defining edges as single events.

at a periodical level (or on a subset of papers [81, 98, 79, 86]), meaning that individual papers alongside their citation statistics are unified under the specific periodical they were published. Such an aggregation has some merits, such as scalability efficiency ($\#\text{papers} \gg \#\text{periodicals}$) but leads to information loss at the paper level, especially for interdisciplinary journals [8, 55].

Various prominent Graph Representation Learning (GRL) methods have also been applied to citation graphs [82, 52] as they have been very popular network choices for assessing downstream task performance, such as link prediction and node classification [70, 31, 73]. Recently, Graph Neural Networks (GNN)s have also been used including GraphSAGE [34], the Adaptive Channel Mixing GNN [57] and Convolutional Graph Neural Networks [47]. Despite these works defining strong models, powerful link predictors, and node classifiers they do not explicitly account for impact characterization, nor for the dynamic way that paper citations appear.

Notably, citation networks are dynamic. Whereas dynamic modeling approaches can uncover structures obscured when aggregating networks across time to form static networks, the dynamic modeling approaches are in general based on the assumption that multiple links occur between the dyads in time. Importantly, for continuous-time modeling this has typically been accounted for using Poisson Process likelihoods [5, 21, 10] including likelihoods accounting for burstiness and reciprocating behaviors by use of the Hawkes process [5, 2, 15, 108, 21]. To account for the high degree of complex interactions in time, advanced dynamic latent representations have further been proposed considering both discrete-time [43, 37, 36, 70, 31, 18, 19, 45, 80, 83] and continuous-time dynamics [5, 2, 15, 21, 10], including GNNs with time-evolving latent representations [92, 78]. For surveys of such dynamic modeling approaches see also [105, 44].

Importantly, citation networks are a class of dynamic networks characterized by a single event occurring between dyads, which we denote as Single-Event Networks (SEN). I.e., links occur only once at the time of the paper publication. However, neither of the existing dynamic network modeling approaches explicitly account for SENs. Whereas continuous-time modeling approaches are designed for multiple events, thereby easily over-parameterizing such highly sparse networks, static networks can easily be applied to such networks by disregarding the temporal structure but thereby potentially miss important structural information given by the event time. Despite these limitations, to the best of our knowledge, existing dynamic network modeling approaches do not explicitly account for single-event occurrences. In Figure 1, we provide an example of three cases of networks that define static, traditional event-based dynamic networks, as well as SENs. We here observe how static networks are completely blind to the temporal information that single-event networks capture while it is also evident that they differ from traditional event-based temporal networks where each dyad can have multiple events across time.

When modeling SENs, the single event occurrence makes the networks highly sparse. To account for the high degree of sparsity of SENs we use as a starting point the static Latent Distance Modeling (LDM) approaches proposed in [40] in which static networks are embedded in a low dimensional space and the relative distance between the nodes used to parameterize the probability of observing

links between the nodes. Importantly, these modeling approaches have been found to provide easily interpretable low-dimensional ($D = 2$ and $D = 3$) network representations with favorable representation learning performance in tasks including link prediction and node classification [61, 59]. The LDM has been generalized to distances beyond Euclidean, including squared Euclidean distances and hyperbolic embeddings [66, 67] as well as to account for degree heterogeneity through the use of node-specific biases (denoted random effects) [39, 48, 61] which we presently refer to as the mass of a paper. Notably, we define paper masses based on their citation dynamics through time, regulated by their distance in a latent space used to embed the structure of the citation network. Specifically, to account for single-event network dynamics, we endow the cited papers (receiving nodes) with a temporal profile in which a parametric function as used for traditional paper impact assessment [77] is employed to regulate the nodes' citation activity in time forming the *Dynamic Impact Single-Event Embedding Model* (DISEE). In particular, our contributions are

- **We derive the single-event Poisson Process (SE-PP).**
  As paper citation networks only include a single event we augment the Poisson Process likelihood to have support only for single events forming the single event Poisson Process.
- **We propose the DISEE model based on the SE-PP for SENs.**
  We characterize the rate of interaction within a latent distance model augmented such that citations are generated relative to the degree to which a paper cites and a paper is being cited at a given time point interpreted as masses of the citing and cited papers in which the mass of the cited paper is dynamically evolving.
- **We demonstrate how DISEE reconciles conventional impact modeling with latent distance embedding procedures.**
  We demonstrate how DISEE enables accurate dynamic characterization of citation impact similar to conventional paper impact modeling procedures while at the same time providing low-dimensional embeddings accounting for the structure of citation networks. We highlight this reconciliation on three real networks covering three distinct fields of science.

The paper is organized as follows. In Section 2, we present the single-event Poisson Process (SE-PP) for the modeling of single-event networks (SENs). In Section 4, we demonstrate how existing embedding procedures can be reconciled with dynamic impact modeling using the SE-PP by the proposed Dynamic Impact Single-Event Embedding Model (DISEE). In Section 5, we present our results on the three distinct citation networks contrasting the performance to the corresponding conventional impact dynamic modeling, as well as, the powerful static LDM [61]. Section 6 concludes our results and points to future directions of research.

2 The Single-Event Poisson Process

Many real networks continually change over time, with new nodes and connections emerging as the network evolves. Prominent examples of such networks include citation networks, user-item review and rating graphs, collaboration graphs, and contact networks. Contact networks and collaboration networks typically include multiple events between the dyads. However, for user-item review and rating networks, an individual's activity history forms new connections between the online profile and the reviewed/rated products. It can also be argued that the same link typically does not occur multiple times as a person once reviewing a product does not create additional reviews of the same product. Importantly, citation networks are characterized by node pairs (i.e., dyads) that can have only one event defined by the point in time at which a citing paper cites another paper. In this regard, we assume that the observed networks are composed maximally of single-event node pairs (dyads), which we call *Singe-Event-Networks* (SENs) and once an event between two nodes has occurred no more event are admissible between these two nodes, see also Figure 1 *Right panel*.

Before presenting our modeling strategy for the links of networks, we will first establish the notations used throughout the paper. We utilize the conventional symbol, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, to denote a directed Single-Event-Network over the timeline $[0, T]$ where $\mathcal{V} = \{1, \dots, N\}$ is the vertex and $\mathcal{E} \subseteq \mathcal{V}^2 \times [0, T]$ is the edge set such that each node pair has at most one link. Hence, a tuple, $(i, j, t_{ij}) \in \mathcal{E}$, shows a directed event (i.e., instantaneous link) from source node j to target i at time $t_{ij} \in [T]$, and there can be at most one (i, j, t_{ij}) element for each $(i, j) \in \sqsubseteq^{\in}$ and some $t_{ij} \in [0, T]$.

We always assume that the timeline starts at 0 and the last time point is T, and we represent the interval by symbol, $[T]$. We employ $t_1 \leq t_2 \leq \cdots \leq t_N$ to indicate the appearance times of the

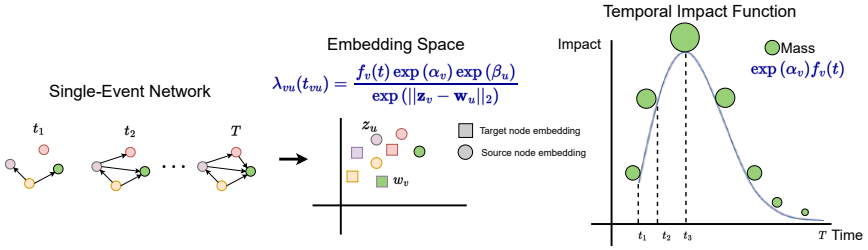


Figure 2: DISEE procedure overview. Given a Single-Event Network (SEN) as an input, the model defines an intensity function introducing two sets of static embeddings distinguishing between source $\mathbf{w}_u$ and target $\mathbf{z}_v$ node embeddings. Furthermore, each node is assigned its own random effect, distinguishing again the source β_u and target α_v roles. The random effects can be parameterized to represent source and target masses through the exponential function. Finally, for each target node of the network, the model further defines an impact function $f_v(t)$ yielding a temporal impact characterization of the nodes' incoming link dynamics which controls the nodes' mass in time as, $\exp(\alpha_v) f_v(t)$.

corresponding nodes $1, 2, \ldots, N \in \mathcal{V}$, and we suppose that node labels are sorted with respect to their incoming edge times. In other words, if $i < j$, then we know that there is a node $k \in \mathcal{V}$ such that $t_{ik} \leq t_{jl}$ for all $l \in \mathcal{V}$.

The Inhomogeneous Poisson Point (IPP) process is a widely employed approach for modeling the number of events exhibiting varying characteristics depending on the time they occur [30]. They are parametrized by an *intensity* or *rate function* representing the average event density, and the probability of sampling m event points on the interval $[T]$ is given by

$$p_M(M(T) = m) := \frac{[\Lambda(0,T)]^m}{m!} \exp(-\Lambda(0,T)), \tag{1}$$

where $M(T)$ is the random variable showing the number of events occurring over the interval $[T]$, and $\Lambda(T) := \int_0^T \lambda(t')\mathrm{d}t'$ for the intensity function $\lambda : [T] \to \mathbb{R}^+$. We refer unfamiliar readers to the work [90] for more details concerning the process. It is worth noting here that earlier studies [101, 61] have demonstrated that adopting the Poisson likelihood for modeling binary relationships does not degrade the methods' predictive performance and its ability to uncover the network structure.

In this regard, we employ a Poisson point process for characterizing the occurrence time of a link (i.e., a single event point indicating the publication or citation time), unlike their conventional practice in modeling the occurrence of an arbitrary number of events between a pair of nodes. Hence, we suppose that a pair can have at most one interaction (i.e., link), and we discretize the probability of sampling m events given in Equation (1) as having an event and no event cases. More formally, by applying Bayes' rule, we can write it as a conditional distribution of $M(t)$ being equal to $m \in \{0, 1\}$ as follows:

$$\begin{aligned} p_{M|M\leq 1}(M(T) = m) &= \frac{p_{M,M\leq 1}(M(T) = m, M(T) \leq 1)}{p_{M\leq 1}(M(T) \leq 1)} = \frac{p_M(M(T) = m)}{p_M(M(T) = 0) + p_M(M(T) = 1)} \\ &= \frac{\exp(-\Lambda(T))[\Lambda(T)]^m}{\exp(-\Lambda(T)) + \exp(-\Lambda(T))\Lambda(T)} \end{aligned} \tag{2}$$

Therefore, the conditional probability of having an event for the proposed *Single-Event Poisson Process* is equal to:

$$p_{M|M\leq 1}(M(T) = 1) = \frac{\Lambda(T)}{1 + \Lambda(T)}. \tag{3}$$

It is also not difficult to derive the likelihood function of the process based on Eq (3). Let (Y, Θ) be random variables where Y shows whether a link exists and Θ indicates the time of the corresponding

link (if it exists). Then, we can write the likelihood of (Y, Θ) evaluated at $(1, t^*)$ as follows:

$$p_{Y,\Theta}\left(1, t^*\right) = p_Y\left\{Y = 1\right\} p_{\Theta|Y}\left\{\Theta = t^* | Y = 1\right\} = \left(\frac{\Lambda(T)}{1 + \Lambda(T)}\right)\left(\frac{\lambda(t^*)}{\Lambda(T)}\right) = \frac{\lambda(t^*)}{1 + \Lambda(T)} \quad (4)$$

As a result, we can write the log-likelihood of the whole network by assuming that each dyad follows the Single-Event Poisson Process as follows:

$$\mathcal{L}_{SE-PP}(\Omega) := \log p(\mathcal{G}|\Omega) = \sum_{1 \leq i,j \leq N} \Big(y_{ij} \log \lambda(t_{ij}) - \log\left(1 + \Lambda_{ij}(t_i, T)\right)\Big) \quad (5)$$

where Ω is the model hyper-parameters and $\Lambda_{ij}(t_i, T) := \int_{t_i}^{T} \lambda_{ij}(t')\mathrm{d}t'$. Note that for a homogeneous Poisson process with constant intensity λ_{ij} for each node pair i and j, the probability of having an event throughout the timeline is equal to $\Lambda_{ij}(T)/(1 + \Lambda_{ij}(T)) = T\lambda_{ij}/(1 + T\lambda_{ij})$ by Equation (3). In this regard, the objective function stated in Equation (5) is equivalent to a static Bernoulli model [41]:

$$\mathcal{L}_{Bern}(\Omega) := \log p(\mathcal{G}|\Omega) = \sum_{i,j \in \mathcal{V}} \Big(y_{ij} \log\left(\tilde{\lambda}_{ij}\right) - \log\Big(1 + \tilde{\lambda}_{ij}\Big)\Big), \quad (6)$$

where we have used the re-parameterization $T\lambda_{ij} = \tilde{\lambda}_{ij}$.

3 Dynamic Impact Characterization

In the realm of impact analysis and risk assessment, characterizing dynamic events is pivotal in understanding and managing potential consequences. We know that papers generally undergo the process of aging over time since novel works introduce more original concepts. In this regard, we model the distribution of the impact of a paper $\{i\}$ by the TRUNCATED normal distribution:

$$f_i(t) = \frac{1}{\sigma} \frac{\phi(\frac{t-\mu}{\sigma})}{\Phi(\frac{\kappa-\mu}{\sigma}) - \Phi(\frac{\rho-\mu}{\sigma})} \quad (7)$$

where μ and σ are the parameters of the distribution which lie in $(\rho, \kappa) \in \mathbb{R}$, $\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right)$, and $\Phi(\cdot)$ is the cumulative distribution function $\Phi(x) = \frac{1}{2}\Big(1 + \mathrm{erf}(\frac{x}{\sqrt{2}})\Big)$. In addition, as an alternative impact function, and similar to [98], we consider the LOG-NORMAL distribution:

$$f_i(t) = \frac{1}{t\sigma\sqrt{2\pi}} \exp\left(-\frac{\ln\left(t - \mu\right)^2}{2\sigma^2}\right) \quad (8)$$

where μ and σ are the parameters of the distribution. Such distributions are particularly valuable for capturing the inherent variability and asymmetry in the lifecycle of a paper.

4 Single-Event Network Embedding by the Latent Distance Model

Our main purpose is to represent every node of a given single-event network in a low D-dimensional latent space ($D \ll N$) in which the pairwise distances in the embedding space should reflect various structural properties of the network, like homophily and transitivity [61]. For instance, in the *Latent Distance Model* [41], one of the pioneering works, the probability of a link between a pair of nodes depended on the log-odds expression, γ_{ij}, as $\alpha - \|\mathbf{z}_i - \mathbf{z}_j\|_2$ where $\{\mathbf{z}_i\}_{i \in \mathcal{V}}$ are the node embeddings, and $\alpha \in \mathbb{R}$ is the global bias term responsible for capturing the global information in the network. It has been proposed for undirected graphs but can be extended for directed networks as well by simply introducing another node representation vector $\{\mathbf{w}_i\}_{i \in \mathcal{V}}$ in order to differentiate the roles of the node as source (i.e., sender) and target (i.e., receiver). By the further inclusion of two sets of random effects $\{\alpha_i, \beta_j\}$ describing the in and out degree heterogeneity, respectively, we can define the log-odds(Bernoulli) and log-rate (Poisson) [61] expression as:

$$\gamma_{ij} = \alpha_i + \beta_j - \|\mathbf{z}_i - \mathbf{w}_j\|_2 \quad (9)$$

We can now combine a dynamic impact characterization function with the *Latent Distance Model*, to obtain an expression for the intensity function of the proposed *Single-Event Poisson Process*, as:

$$\lambda_{ij}(t_{ij}) = \frac{f_i(t_{ij}) \exp(\alpha_i) \exp(\beta_j)}{\exp(\|\mathbf{z}_i - \mathbf{w}_j\|_2)}. \tag{10}$$

Combining the intensity function of Equation (10) with the log-likelihood expression of Equation (5) yields the *Dynamic Impact Single-Event Embedding Model* (DISEE). Under such a formulation, we exploit the time information data indicating when links occur through time, so we can grasp a more detailed understanding of the evolution of networks, generate enriched node representations, and quantify a node's temporal impact on the network.

4.1 Case-Control Inference

With DISEE being a distance model, it scales prohibitively as $\mathcal{O}(N^2)$ since the all-pairs distance matrix needs to be calculated. In order to scale the analysis to large-scale networks we adopt an unbiased estimation of the log-likelihood similar to a case-control approach [75]. In our formulation, we calculate the log-likelihood as:

$$\begin{aligned}\log p_{ij}(\mathcal{G}|\Omega) = &\sum_{j:y_{ij}=1} \left(y_{ij} \log\left(\lambda_{ij}(t^*_{ij})\right) - \log\left(1 + \int_{t_i}^{T} \lambda_{ij}(t')dt'\right)\right) \\ &+ \sum_{j:y_{ij}=0} - \log\left(1 + \int_{t_i}^{T} \lambda_{ij}(t')dt'\right) \\ &= l_1 + l_0 \end{aligned} \tag{11}$$

Large networks are usually sparse so the link (case) likelihood contribution term l_1 can be calculated analytically, even for massive networks. The non-link (control) likelihood contribution term l_0 has a quadratic complexity $\mathcal{O}(N^2)$ in terms of the size of the network, and thus its computation is infeasible. For that, we introduce an unbiased estimator for $l_{i,0}$ which is regarded as a population total statistic [75]. We estimate the non-link contribution of a node $\{i\}$ via:

$$l_{i,0} = \frac{N_{i,0}}{n_{i,0}} \sum_{k=1}^{n_{i,0}} - \log\left(1 + \int_{t_i}^{T} \lambda_{ik}(t')dt'\right), \tag{12}$$

where $N_{i,0}$ is the number of total non-links (controls) for node $\{i\}$, and $n_{i,0}$ is the number of samples to be used for the estimation. We set the number of samples based on the node degrees as $n_{i,0} = 5 * \text{degree}_i$. This makes inference scalable defining an $\mathcal{O}(cE)$ space and time complexity.

4.2 Model ablations

We define an Impact Function Model (IFM), where only the impact function is fitted to the target nodes (cited papers) describing their link (citation) dynamics. Comparing with IFM will allow us to validate the quality of the impact characterization of DISEE. We further contrast our model to a Preferential Attachment Model (PAM) setting where the embedding dimension is set as $D = 0$, providing a quantification of the importance of including an impact function and an embedding space in DISEE. In addition, we consider a combination of an Impact Function Model with a Preferential Attachment Model, defining a Temporal Preferential Attachment Model (TPAM). Compared with the TPAM we aim to verify the importance of introducing an embedding space characterization in citation networks. Finally, we systematically contrast the performance of DISEE to conventional static latent distance modeling (LDM) corresponding to setting the impact function to be constant $f_i(t) \propto 1$ in DISEE. The LDM is a very powerful link predictor [60, 59] and contrasting its performance against DISEE will help us showcase the successful reconciliation of static latent space network embedding approaches with classical dynamic impact assessments of citation networks. In Table 1, we provide the rate formulation of each of the considered model ablations and the corresponding model characteristics in terms of impact characterization, definition of an embedding space, and link prediction.

Table 1: DISEE model and multiple model ablations. Specifically, we consider 1) An Impact Function Model (IFM) which characterizes only the impact based on the incoming citation dynamics of each paper. 2) A Preferential Attachment Model (PAM) which defines citing and cited masses, yielding essentially a degree-based model. 3) A combination of the Impact Function Model and the Preferential Attachment Model defining a Temporal Impact Function Model (TPAM). 4) A bipartite formulation of the classic Latent Distance Model (LDM) [41]. For each model, we provide the rate formulation, as well as, its capacity in terms of impact characterization, definition of an embedding space, and link prediction capability.

Model name	**Rate formulation**	**Impact**	**Embedding space**	**Link prediction**
IFM	$f_i(t)$	✓	✗	✗
PAM	$\exp(\alpha_i)\exp(\beta_j)$	✗	✗	✓
TPAM	$f_i(t)\exp(\alpha_i)\exp(\beta_j)$	✓	✗	✓
LDM	$\frac{\exp(\alpha_i)\exp(\beta_j)}{\exp(\|\|\boldsymbol{z}_i-\boldsymbol{w}_j\|\|_2)}$	✗	✓	✓
DISEE	$\frac{f_i(t)\exp(\alpha_i)\exp(\beta_j)}{\exp(\|\|\boldsymbol{z}_i-\boldsymbol{w}_j\|\|_2)}$	✓	✓	✓

Table 2: AUC-ROC scores for varying representation sizes over three citation networks.

	ML		*Phys*		*SoSci*	
Dimension (D)	2	3	2	3	2	3
PAM	0.810		0.838		0.796	
TPAM TRUNCATED	0.806		0.836		0.790	
TPAM LOG-NORMAL	0.814		0.839		0.799	
LDM	0.969	0.976	0.963	0.973	0.956	0.963
DISEE TRUNCATED	0.968	0.977	0.962	0.973	0.960	0.965
DISEE LOG-NORMAL	0.969	0.976	0.961	0.970	0.957	0.964

5 Results and Discussion

In this section, we will evaluate how successfully DISEE reconciles traditional impact quantification approaches with latent distance modeling. Specifically, we test the proposed approach's effectiveness in the link prediction task by comparing it to the classical LDM which is not time-aware and does not quantify temporal impact. We also consider multiple model ablations that are either able to characterize a node's impact or account for GRL, i.e. define node embeddings, but not both. For the task of link prediction, we remove 20% of network links and we sample an equal amount of non-edges as negative samples and construct the test set. Notably, these negative samples are sampled in a time-aware manner, meaning that we consider only pairs that are possibly to exist as missing links in the network (i.e. we do not consider node pairs where missing citations refer to papers citing future papers, as the target paper did not exist the time when the source paper was published). The link removal is designed in such a way that the residual network stays connected. Analytically, for each network, we do not consider removing links that make up the minimum spanning tree of the graph. For the evaluation, we consider both the Receiver Operator Characteristic and Precision-Recall Area Under Curve scores, as these are metrics not sensitive to the class imbalance between links and non-links. We then continue by evaluating the quality of impact expression of DISEE by visually presenting the inferred impact functions and comparing them against the IFM model. Finally, we visualize the model's learned temporal space representing the target papers, accounting for their temporal impact in terms of their mass at a specific time point, and characterizing the different papers' lifespans.

Table 3: AUC-PR scores for varying representation sizes over three citation networks.

	ML		*Phys*		*SoSci*	
Dimension (D)	2	3	2	3	2	3
PAM	0.823		0.841		0.814	
TPAM TRUNCATED	0.812		0.836		0.806	
TPAM LOG-NORMAL	0.818		0.837		0.812	
LDM	0.971	0.977	0.965	0.974	0.961	0.967
DISEE TRUNCATED	0.970	0.977	0.964	0.973	0.963	0.969
DISEE LOG-NORMAL	0.971	0.977	0963	0.972	0.962	0.968

Table 4: Statistics of networks. $|\mathcal{V}_1|$: Number of target nodes, $|\mathcal{V}_2|$: Number of source nodes, $|\mathcal{E}|$: Total number of links.

	$\lvert\mathcal{V}_1\rvert$	$\lvert\mathcal{V}_2\rvert$	$\lvert\mathcal{E}\rvert$
Machine Learning	22,540	148,703	526,226
Physics	20,012	51,996	573,378
Social Science	12,930	100,402	288,012

5.1 Datasets

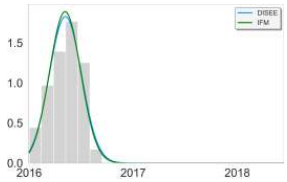
In our experiments, we employ three real citation networks. Specifically, we use the OpenAlex dataset [72], exploring highly impactful scientific domains such as (i) *Machine Learning*, (ii) *Physics*, and (iii) *Social Science*. In order to be able to characterize scientific impact, we initially consider papers that have been cited at least ten times, defining three directed networks. Since there is no guarantee that such a filtering approach will define a network made up of nodes with a minimum degree of ten citations, we define a zero mass for the papers (target nodes) that survive the initial thresholding and have less than ten citations. This yields a directed bipartite structure where target nodes have at least ten citations. Analytically, the network statistics are given in Table 4.

5.2 Link prediction

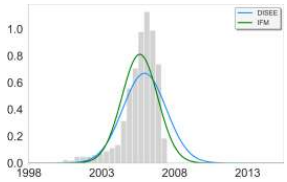
For the link prediction experiments, and for each network, we remove 20% of the edges which are not the edges that construct the minimum spanning tree. Consequently, the residual network is guaranteed to stay connected. The removed edges are combined with a sample of the same number of node pairs, that are not the edges of the original network, to construct the negative instances for the testing set. We utilize the residual network to learn the node embeddings used for the link prediction experiments. We compare the results of DISEE with the introduced model ablations in terms of the Area Under Curve-Receiver Operating Characteristic (AUC-ROC), and the Area Under Curve-Precision Recall (AUC-PR) scores, as presented in Table 2 and Table 3, respectively. Scores are presented as the mean value of three independent runs of the Adam optimizer [46] (error bars were found in the 10^{-4} scale and thus omitted). We here observe, that the best performance is achieved by model specifications that define an embedding space, i.e. the DISEE and LDM models. The Preferential Attachment Models in both the static (PAM) and temporal (TPAM) versions are characterized by an approximately 15% decrease in their link prediction score in both AUC-ROC and AUC-PR. This highlights the importance and benefits of the predictive performance an embedding space provides. Contrasting now, the performance of DISEE against the LDM, we witness almost identical scores, verifying that DISEE successfully inherited the link prediction power of the LDM. Comparing the two distribution choices for the impact function (TRUNCATED NORMAL and LOG-NORMAL) we again observe very similar scores.

5.3 Impact quantification

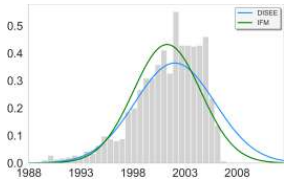
We now continue by addressing the quality of paper impact characterization based on a target paper's incoming citation dynamics. In Figure 3 and Figure 4, we provide inferred impact functions of the



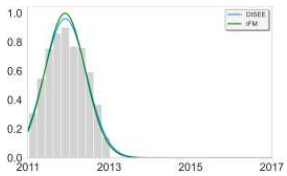
(a) Deep Residual Learning for Image Recognition [35].



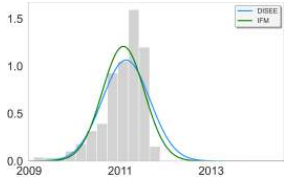
(b) Gradient-based learning applied to document recognition [51].



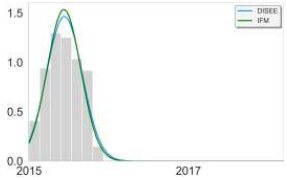
(c) Structural equation modeling in practice [1].



(d) LIBSVM: A library for support vector machines [11].

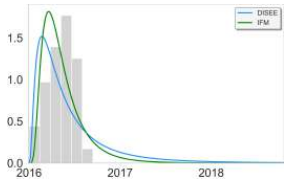


(e) ImageNet: A large-scale hierarchical image database [16].

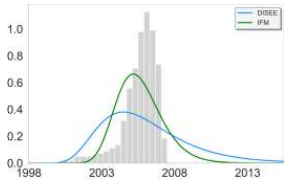


(f) Going deeper with convolutions [91].

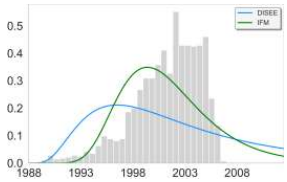
Figure 3: *Machine Learning*: DISEE TRUNCATED and IFM TRUNCTATED models inferred impact function visualizations compared to the true citation histogram, for six popular *Machine Learning* papers with different citation dynamics.



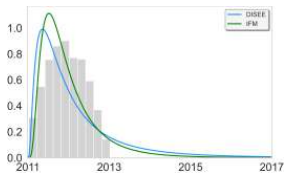
(a) Deep Residual Learning for Image Recognition.



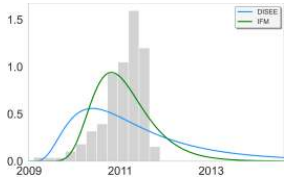
(b) Gradient-based learning applied to document recognition.



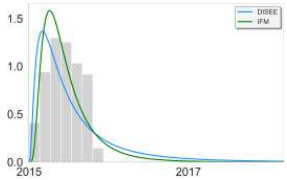
(c) Structural equation modeling in practice.



(d) LIBSVM: A library for support vector machines.



(e) ImageNet: A large-scale hierarchical image database.



(f) Going deeper with convolutions.

Figure 4: *Machine Learning*: DISEE LOG-NORMAL and IFM LOG-NORMAL models inferred impact function visualizations compared to the true citation histogram, for six popular *Machine Learning* papers with different citation dynamics.

DISEE and IFM, under the TRUNCATED normal and LOG-NORMAL distributions, respectively. We further show the true impact dynamics through the citation histogram for each one of the corresponding papers. For the TRUNCATED case, we observe that DISEE and IFM provide very similar (and in some cases identical) impact functions that well capture the underlying citation histogram. In the case of the LOG-NORMAL distribution, we witness an agreement between DISEE and IFM models when the paper lifespan does not exceed 2 years. For larger lifespans DISEE defines a larger standard deviation than the IFM returning much heavier tails. Both models when compared to the true citation histogram provide much heavier tails when the paper lifespan exceeds the 2-year threshold. The LOG-NORMAL distribution is not invariant to the scale of the x-axis (contrary to the TRUNCATED normal which is scale-invariant) and this can be potentially a reason for observing

this kind of behavior, meaning that the choice of the time resolution is not optimal. Nevertheless, the TRUNCATED normal distribution seems to very accurately represent the true citation dynamics, defining correct distribution tails, but in some cases, the LOG-NORMAL heavier tails may be more appropriate for future impact predictions (as papers stay "alive" longer).

5.4 Space visualization

Finally, we here provide embedding space visualizations of the target papers, accounting for their temporal impact in terms of their mass at a specific time point. Analytically, Figure 6 shows the evolution of the embedding space for the domain of *Machine Learning* from the year 1988 until 2018. We here observe how papers are published in a specific year and after they accumulate a specific amount of impact/mass they perish in the next years/snapshots of the network. It is also worth mentioning, that the domain of *Machine Learning* has undergone a significant increase in the paper outputs as the years progress. As the years progress, paper masses reach much larger magnitudes than in the earlier years, defining higher research significance, and accumulating higher citation numbers and impact which can be explained by the increase in published *Machine Learning* works. Figure 5 shows the present embedding space image of the domain of *Machine Learning*. It is evident that many papers stay active throughout the years but with decreasing masses. It is surprising to see that the papers with the largest masses are relatively old papers from the early 2000s with a few cases published in more recent years around 2010s.

6 Conclusion

We have proposed the Dynamic Impact Single-Event Embedding Model (DISEE), a reconciliation between traditional impact quantification approaches with a Latent Distance Model (LDM). We have focused on Single-Event Networks (SENs), and more specifically in citation networks, where we to the best of our knowledge for the first time derived and explored the Single-Event Poisson Process (SE-PP). Such a process defines an appropriate likelihood allowing for a principled analysis of single-events networks. In order to define powerful ultra-low dimensional network embeddings, we turn to the representation power of the directed network version of the LDM. Specifically, for every paper we define static embeddings distinguishing between source and target roles, i.e. we introduced a different position in the latent space for the roles of papers when citing or being cited. In addition, we defined paper random effects that can be reparametrized to represent paper masses, again distinguishing between "being cited" and "citing" masses. For the "being cited" mass, we introduced a temporal impact function that characterized the incoming citation dynamics. eligible for impact quantification. The impact function is parameterized through appropriate probability density functions, including the log-normal, as well as, the truncated normal distributions. Through extensive experiments, we showed that the DISEE had the same link prediction performance as the powerful LDM. Furthermore, we showed that the temporal impact characterization was validated by an Impact Function Model (IFM). These results showcase that the DISEE model successfully reconciles powerful embedding approaches with citation dynamics impact characterization. Finally, visualizations of the embedding space for target papers provided accurate representations that described the birth and death of papers following their impact lifespans as years pass and science moves forward.

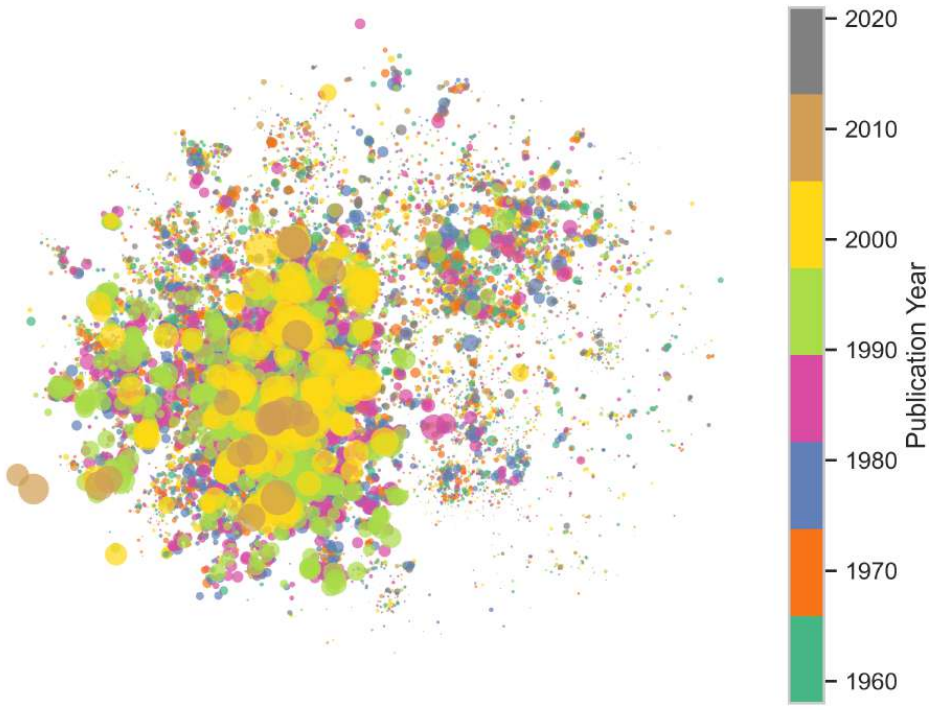


Figure 5: *Machine Learning*: DISEE TRUNCATED embedding space visualization for all target papers published before the year 2023. Node sizes are based on each paper's current mass, $f_i(t) * \exp(\alpha_i)$, and thus papers with zero mass are not visible denoting the end of their scientific relevance or "lifespan". Nodes are color-coded based on their publication year.

(a) 1988 (b) 1998 (c) 2008 (d) 2018

Figure 6: *Machine Learning*: DISEE TRUNCATED embedding space evolution throughout the years. Node sizes are based on each paper's mass, $f_i(t) \exp(\alpha_i)$, showcasing how papers reach the end of their scientific relevance or "lifespan" by disappearing from the embedding space as time progresses. Nodes are color-coded based on their publication year.

References

[1] James C. Anderson and David W. Gerbing. Structural equation modeling in practice: A review and recommended two-step approach. *Psychological bulletin*, 103(3):411–423, May 1988.

[2] Makan Arastuie, Subhadeep Paul, and Kevin Xu. CHIP: A hawkes process model for continuous-time networks with scalable and consistent estimation. In *NeurIPS*, volume 33, pages 16983–16996, 2020.

[3] Albert-László Barabási, Chaoming Song, and Dashun Wang. Handful of papers dominates citation. *Nature*, 491(7422):40–41, 2012.

[4] Harish S. Bhat, Li-Hsuan Huang, Sebastian Rodriguez, Rick Dale, and Evan Heit. Citation prediction using diverse features. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, pages 589–596, 2015.

[5] Charles Blundell, Jeff Beck, and Katherine A Heller. Modelling reciprocating relationships with hawkes processes. In *NeurIPS*, volume 25, 2012.

[6] Johan Bollen, Herbert Van de Sompel, Aric Hagberg, Luis Bettencourt, Ryan Chute, Marko A. Rodriguez, and Lyudmila Balakireva. Clickstream data yields high-resolution maps of science. *PLOS ONE*, 4(3):1–11, 03 2009.

[7] George J Borjas and Kirk B Doran. Which peers matter? the relative impacts of collaborators, colleagues, and competitors. *Review of economics and statistics*, 97(5):1104–1117, 2015.

[8] Kevin Boyack, Richard Klavans, and Katy Borner. Mapping the backbone of science. *Scientometrics*, 64:351–374, 03 2005.

[9] Katy Börner, Richard Klavans, Michael Patek, Angela M. Zoss, Joseph R. Biberstine, Robert P. Light, Vincent Larivière, and Kevin W. Boyack. Design and update of a classification system: The ucsd map of science. *PLOS ONE*, 7(7):1–10, 07 2012.

[10] Abdulkadir Çelikkanat, Nikolaos Nakis, and Morten Mørup. Piecewise-velocity model for learning continuous-time dynamic node representations. *arXiv preprint arXiv:2212.12345*, 2022.

[11] Chih-Chung Chang and Chih-Jen Lin. Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2(3), may 2011.

[12] Murat Cokol, Ivan Iossifov, Chani Weinreb, and Andrey Rzhetsky. Emergent behavior of growing knowledge about molecular interactions. *Nature biotechnology*, 23:1243–7, 11 2005.

[13] National Research Council et al. Enhancing the effectiveness of team science. 2015.

[14] Feruz Davletov, Ali Selman Aydin, and Ali Cakmak. High impact academic paper prediction using temporal and topological features. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, page 491–498, New York, NY, USA, 2014. Association for Computing Machinery.

[15] Sylvain Delattre, Nicolas Fournier, and Marc Hoffmann. Hawkes processes on large networks. *The Annals of Applied Probability*, 26(1):216 – 261, 2016.

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[17] Jordi Duch, Xiao Han T Zeng, Marta Sales-Pardo, Filippo Radicchi, Shayna Otis, Teresa K Woodruff, and Luís A Nunes Amaral. The possible role of resource requirements and academic career-choice risk on gender differences in publication rate and impact. *PloS one*, 7(12):e51332, 2012.

[18] Daniele Durante and David Dunson. Bayesian Logistic Gaussian Process Models for Dynamic Networks. In *AISTATS*, volume 33, pages 194–201, 2014.

[19] Daniele Durante and David B Dunson. Locally adaptive dynamic networks. *The Annals of Applied Statistics*, 10(4):2203–2232, 2016.

[20] Young-Ho Eom and Santo Fortunato. Characterizing and modeling citation dynamics. *PLOS ONE*, 6(9):1–7, 09 2011.

[21] Xuhui Fan, Bin Li, Feng Zhou, and Scott SIsson. Continuous-time edge modelling using non-parametric point processes. *NeurIPS*, 34:2319–2330, 2021.

[22] Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3):75–174, 2010.

[23] Santo Fortunato, Carl T. Bergstrom, Katy Börner, James A. Evans, Dirk Helbing, Staša Milojević, Alexander M. Petersen, Filippo Radicchi, Roberta Sinatra, Brian Uzzi, Alessandro Vespignani, Ludo Waltman, Dashun Wang, and Albert-László Barabási. Science of science. *Science*, 359(6379):eaao0185, 2018.

[24] Jacob G. Foster, Andrey Rzhetsky, and James A. Evans. Tradition and innovation in scientists' research strategies. *American Sociological Review*, 80(5):875–908, 2015.

[25] Richard Freeman, Eric Weinstein, Elizabeth Marincola, Janet Rosenbaum, and Frank Solomon. Competition and careers in biosciences, 2001.

[26] Eugene Garfield. Citation analysis as a tool in journal evaluation: Journals can be ranked by frequency and impact of citations for science policy studies. *Science*, 178(4060):471–479, 1972.

[27] Lise Getoor and Ben Taskar. *Introduction to Statistical Relational Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2007.

[28] Michel L. Goldstein, Steven A. Morris, and Gary G. Yen. Group-based yule model for bipartite author-paper networks. *Phys. Rev. E*, 71:026108, Feb 2005.

[29] Michael Golosovsky and Sorin Solomon. Runaway events dominate the heavy tail of citation distributions. *The European Physical Journal Special Topics*, 205(1):303–311, 2012.

[30] Jonatan A. González, Francisco J. Rodríguez-Cortés, Ottmar Cronie, and Jorge Mateu. Spatio-temporal point process statistics: A review. *Spatial Statistics*, 18:505–544, 2016.

[31] Aditya Grover and Jure Leskovec. Node2Vec: Scalable feature learning for networks. In *KDD*, pages 855–864, 2016.

[32] S. Gualdi, M. Medo, and Y.-C. Zhang. Influence, originality and similarity in directed acyclic graphs. *Europhysics Letters*, 96(1):18004, sep 2011.

[33] S. Gualdi, C. H. Yeung, and Y.-C. Zhang. Tracing the evolution of physics on the backbone of citation networks. *Phys. Rev. E*, 84:046104, Oct 2011.

[34] William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *NIPS*, 2017.

[35] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[36] Creighton Heaukulani and Zoubin Ghahramani. Dynamic probabilistic models for latent feature propagation in social networks. In *ICML*, pages 275–283, 2013.

[37] Tue Herlau, Morten Mørup, and Mikkel Schmidt. Modeling temporal evolution and multiscale structure in networks. In *ICML*, pages 960–968, 2013.

[38] Jorge E Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences*, 102(46):16569–16572, 2005.

[39] Peter D Hoff. Bilinear mixed-effects models for dyadic data. *JASA*, 100(469):286–295, 2005.

[40] Peter D Hoff, Adrian E Raftery, and Mark S Handcock. Latent space approaches to social network analysis. *JASA*, 97(460):1090–1098, 2002.

[41] Peter D Hoff, Adrian E Raftery, and Mark S Handcock. Latent space approaches to social network analysis. *JASA*, 97(460):1090–1098, 2002.

[42] Alfonso Ibáñez, Pedro Larranaga, and Concha Bielza. Predicting citation count of bioinformatics papers within four years of publication. *Bioinformatics (Oxford, England)*, 25:3303–9, 10 2009.

[43] Katsuhiko Ishiguro, Tomoharu Iwata, Naonori Ueda, and Joshua Tenenbaum. Dynamic infinite relational model for time-varying relational data analysis. *NeurIPS*, 23, 2010.

[44] Seyed Mehran Kazemi, Rishab Goel, Kshitij Jain, Ivan Kobyzev, Akshay Sethi, Peter Forsyth, and Pascal Poupart. Representation learning for dynamic graphs: A survey. *JMLR*, 21(70):1–73, 2020.

[45] Bomin Kim, Kevin H Lee, Lingzhou Xue, and Xiaoyue Niu. A review of dynamic network models with latent variables. *Statistics surveys*, 12:105, 2018.

[46] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

[47] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks, 2017.

[48] Pavel N. Krivitsky, Mark S. Handcock, Adrian E. Raftery, and Peter D. Hoff. Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models. *Social Networks*, 31(3):204 – 213, 2009.

[49] Vincent Larivière, Yves Gingras, Cassidy R Sugimoto, and Andrew Tsou. Team size matters: Collaboration and scientific impact since 1900. *Journal of the Association for Information Science and Technology*, 66(7):1323–1332, 2015.

[50] Vincent Larivière, Chaoqun Ni, Yves Gingras, Blaise Cronin, and Cassidy R Sugimoto. Bibliometrics: Global gender disparities in science. *Nature*, 504(7479):211–213, 2013.

[51] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[52] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graphs over time: Densification laws, shrinking diameters and possible explanations. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, KDD '05, page 177–187, New York, NY, USA, 2005. Association for Computing Machinery.

[53] Adrian Letchford, Helen Moat, and Tobias Preis. The advantage of short paper titles. *Royal Society Open Science*, 2:150266, 08 2015.

[54] Adrian Letchford, Tobias Preis, and Helen Susannah Moat. The advantage of simple paper abstracts. *Journal of Informetrics*, 10(1):1–8, 2016.

[55] L. Leydesdorff. Various methods for the mapping of science. *Scientometrics*, pages 295 – 324, 1987.

[56] David Liben-Nowell and Jon Kleinberg. The link prediction problem for social networks. In *CIKM*, page 556–559, 2003.

[57] Sitao Luan, Chenqing Hua, Qincheng Lu, Jiaqi Zhu, Mingde Zhao, Shuyuan Zhang, Xiao-Wen Chang, and Doina Precup. Is heterophily a real nightmare for graph neural networks to do node classification?, 2021.

[58] Staša Milojević. Principles of scientific research team formation and evolution. *Proceedings of the National Academy of Sciences*, 111(11):3984–3989, 2014.

[59] Nikolaos Nakis, Abdulkadir Çelikkanat, and Morten Mørup. HM-LDM: A hybrid-membership latent distance model. In *CNA XI*, pages 350–363. Springer International Publishing, 2023.

[60] Nikolaos Nakis, Abdulkadir Çelikkanat, Sune Lehmann Jørgensen, and Morten Mørup. A hierarchical block distance model for ultra low-dimensional graph representations. 2022.

[61] Nikolaos Nakis, Abdulkadir Çelikkanat, Sune Lehmann, and Morten Mørup. A hierarchical block distance model for ultra low-dimensional graph representations. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–14, 2023.

[62] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45(2):167–256, 2003.

[63] Mark EJ Newman. Scientific collaboration networks. i. network construction and fundamental results. *Physical review E*, 64(1):016131, 2001.

[64] Mark EJ Newman. Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. *Physical review E*, 64(1):016132, 2001.

[65] Mark EJ Newman. The structure of scientific collaboration networks. *Proceedings of the national academy of sciences*, 98(2):404–409, 2001.

[66] Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. *Advances in neural information processing systems*, 30, 2017.

[67] Maximillian Nickel and Douwe Kiela. Learning continuous hierarchies in the lorentz model of hyperbolic geometry. In *International Conference on Machine Learning*, pages 3779–3788. PMLR, 2018.

[68] Richard Van Noorden. Interdisciplinary research by the numbers. *Nature*, 525:306–307, 2015.

[69] Hao Peng, Qing Ke, Ceren Budak, Daniel M. Romero, and Yong-Yeol Ahn. Neural embeddings of scholarly periodicals reveal complex disciplinary organizations. *CoRR*, abs/2001.08199, 2020.

[70] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In *KDD*, page 701–710, 2014.

[71] Alexander M Petersen, Massimo Riccaboni, H Eugene Stanley, and Fabio Pammolli. Persistence and uncertainty in the academic career. *Proceedings of the National Academy of Sciences*, 109(14):5213–5218, 2012.

[72] Jason Priem, Heather Piwowar, and Richard Orr. Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts, 2022.

[73] Jiezhong Qiu, Yuxiao Dong, Hao Ma, Jian Li, Kuansan Wang, and Jie Tang. Network embedding as matrix factorization: Unifying DeepWalk, LINE, PTE, and Node2Vec. In *WSDM*, 2018.

[74] Filippo Radicchi, Santo Fortunato, and Claudio Castellano. Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences*, 105(45):17268–17272, 2008.

[75] Adrian E. Raftery, Xiaoyue Niu, Peter D. Hoff, and Ka Yee Yeung. Fast inference for the latent space network model using a case-control approximate likelihood. *Journal of Computational and Graphical Statistics*, 21(4):901–919, 2012.

[76] S. Redner. How popular is your paper? an empirical study of the citation distribution. *The European Physical Journal B*, 4(2):131–134, aug 1998.

[77] Sidney Redner. Citation statistics from 110 years of physical review. *Physics Today*, 58, 06 2005.

[78] Emanuele Rossi, Ben Chamberlain, Fabrizio Frasca, Davide Eynard, Federico Monti, and Michael M. Bronstein. Temporal graph networks for deep learning on dynamic graphs. *ICML 2020 Workshop*, 2020.

[79] Martin Rosvall and Carl T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4):1118–1123, jan 2008.

[80] Purnamrita Sarkar and Andrew Moore. Dynamic social network analysis using latent space models. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *NeurIPS*, volume 18, 2005.

[81] Vedran Sekara, Pierre Deville, Sebastian E. Ahnert, Albert-László Barabási, Roberta Sinatra, and Sune Lehmann. The chaperone effect in scientific publishing. *Proceedings of the National Academy of Sciences*, 115(50):12603–12607, 2018.

[82] Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Gallagher, and Tina Eliassi-Rad. Collective classification in network data. *AI magazine*, 2008.

[83] Daniel K. Sewell and Yuguo Chen. Latent space models for dynamic networks. *JASA*, 110(512):1646–1657, 2015.

[84] Hua-Wei Shen, Dashun Wang, Chaoming Song, and Albert-László Barabási. Modeling and predicting popularity dynamics via reinforced poisson processes. *CoRR*, abs/1401.0778, 2014.

[85] Naoki Shibata, Yuya Kajikawa, and Ichiro Sakata. Link prediction in citation networks. *Journal of the American Society for Information Science and Technology*, 63(1):78–85, 2012.

[86] Roberta Sinatra, Dashun Wang, Pierre Deville, Chaoming Song, and Albert-László Barabási. Quantifying the evolution of individual scientific impact. *Science*, 354(6312):aaf5239, 2016.

[87] Mayank Singh, Vikas Patidar, Suhansanu Kumar, Tanmoy Chakraborty, Animesh Mukherjee, and Pawan Goyal. The role of citation context in predicting long-term citation profiles: An experimental study based on a massive bibliographic text dataset. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, page 1271–1280, New York, NY, USA, 2015. Association for Computing Machinery.

[88] Henry Small. Visualizing science by citation mapping. *JASIS*, 50:799–813, 07 1999.

[89] Paula E. Stephan, Reinhilde Veugelers, and Jian Wang. Reviewers are blinkered by bibliometrics. *Nature*, 544:411 – 412, 2017.

[90] Roy L Streit. *Poisson point processes: imaging, tracking, and sensing*. Springer Science & Business Media, 2010.

[91] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions, 2014.

[92] Rakshit Trivedi, Mehrdad Farajtabar, Prasenjeet Biswal, and Hongyuan Zha. Dyrep: Learning representations over dynamic graphs. In *ICLR*, 2019.

[93] Brian Uzzi, Satyam Mukherjee, Michael Stringer, and Benjamin Jones. Atypical combinations and scientific impact. *Science (New York, N.Y.)*, 342:468–72, 10 2013.

[94] Matthew Wallace, Vincent Larivière, and Yves Gingras. Modeling a century of citation distributions. *Journal of Informetrics*, 3, 11 2008.

[95] Ludo Waltman. A review of the literature on citation impact indicators. *Journal of informetrics*, 10(2):365–391, 2016.

[96] Ludo Waltman. A review of the literature on citation impact indicators. *Journal of Informetrics*, 10(2):365–391, 2016.

[97] Ludo Waltman, Nees Jan van Eck, and Anthony FJ van Raan. Universality of citation distributions revisited. *Journal of the American Society for Information Science and Technology*, 63(1):72–77, 2012.

[98] Dashun Wang, Chaoming Song, and Albert-László Barabási. Quantifying long-term scientific impact. *Science*, 342(6154):127–132, 2013.

[99] Jian Wang, Yajun Mei, and Diana Hicks. Comment on “quantifying long-term scientific impact”. *Science*, 345(6193):149–149, 2014.

[100] Jian Wang, Reinhilde Veugelers, and Paula Stephan. Bias against novelty in science: A cautionary tale for users of bibliometric indicators. *Research Policy*, 46(8):1416–1436, 2017.

[101] David Kofoed Wind and Morten Mørup. Link prediction in weighted networks. In *2012 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, Machine Learning for Signal Processing. IEEE, 2012. 2012 IEEE International Workshop on Machine Learning for Signal Processing, MLSP 2012 ; Conference date: 23-10-2012 Through 26-10-2012.

[102] Lingfei Wu, Dashun Wang, and James A Evans. Large teams have developed science and technology; small teams have disrupted it. *arXiv preprint arXiv:1709.02445*, 2017.

[103] Yan Wu, Tom Z.J. Fu, and Dah Ming Chiu. Generalized preferential attachment considering aging. *Journal of Informetrics*, 8(3):650–658, 2014.

[104] Shuai Xiao, Junchi Yan, Changsheng Li, Bo Jin, Xiangfeng Wang, Xiaokang Yang, Stephen M. Chu, and Hongyuan Zhu. On modeling and predicting individual paper citation count over time. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, page 2676–2682. AAAI Press, 2016.

[105] Guotong Xue, Ming Zhong, Jianxin Li, Jia Chen, Chengshuai Zhai, and Ruochen Kong. Dynamic network embedding survey. *Neurocomputing*, 472:212–223, 2022.

[106] Rui Yan, Jie Tang, Xiaobing Liu, Dongdong Shan, and Xiaoming Li. Citation count prediction: Learning to estimate future citations for literature. pages 1247–1252, 10 2011.

[107] Peng Zhang, Jinliang Wang, Xiaojia Li, Menghui Li, Zengru Di, and Ying Fan. Clustering coefficient and community structure of bipartite networks. *Physica A: Statistical Mechanics and its Applications*, 387(27):6869–6875, 2008.

[108] Yuan Zuo, Guannan Liu, Hao Lin, Jia Guo, Xiaoqian Hu, and Junjie Wu. Embedding temporal network via neighborhood formation. In *KDD*, page 2857–2866, 2018.