On The Topology Of Network Fine Structures

A thesis presented for the degree of

Doctor of Philosophy of Imperial College, London

by

Chuan Wen, Loe

Department of Mathematics

Imperial College

180 Queen's Gate, London SW7 2BZ

OCTOBER 30, 2015

I certify that this thesis, and the research to which it refers, are the product of my own work, and that any ideas or quotations from the work of other people, published or otherwise, are fully acknowledged in accordance with the standard referencing practices of the discipline.

Signed:

Copyright

To my parents, my wife and my in-laws.

Abstract

Multi-relational dynamics are ubiquitous in many complex systems like transportations, social and biological. This thesis studies the two mathematical objects that encapsulate these relationships — multiplexes and interval graphs. The former is the modern outlook in Network Science to generalize the edges in graphs while the latter was popularized during the 1960s in Graph Theory.

Although multiplexes and interval graphs are nearly 50 years apart, their motivations are similar and it is worthwhile to investigate their structural connections and properties. This thesis look into these mathematical objects and presents their connections.

For example we will look at the community structures in multiplexes and learn how unstable the detection algorithms are. This can lead researchers to the wrong conclusions. Thus it is important to get formalism precise and this thesis shows that the complexity of interval graphs is an indicator to the precision. However this measure of complexity is a computational hard problem in Graph Theory and in turn we use a heuristic strategy from Network Science to tackle the problem.

One of the main contributions of this thesis is the compilation of the disparate literature on these mathematical objects. The novelty of this contribution is in using the statistical tools from population biology to deduce the completeness of this thesis's bibliography. It can also be used as a framework for researchers to quantify the comprehensiveness of their preliminary investigations.

From the large body of multiplex research, the thesis focuses on the statistical properties of the projection of multiplexes (the reduction of multi-relational system to a single relationship network). It is important as projection is always used as the baseline for many relevant algorithms and its topology is insightful to understand the dynamics of the system.

Acknowledgments

This thesis is possible with the support by the people in the Centre of Complexity Science, Imperial College. They showed me the intrinsic beauty of Physics and to look at Complexity in a systematic manner. Of course the one who deserves the most credit has to be my supervisor, Prof. Henrik J. Jensen.

Henrik slowly guided me out from my familiar and predictable realm of Computer Science and then *toss* me into the chaotic, unorganized and complex real world with nothing but his whiteboard in room EEE1201 and his thoughts. As I develop my scientific maturity as a PhD student, I appreciate his priority to foster my critical thinking over the results of the experiments.

Most importantly he recognized my strength in interdisciplinary studies and have actively support my pursue to deviate my research direction (if necessarily). This gave me the confidence to travel and to engage with scientists around the world including the most memorable experience in Antarctica.

Finally the list of people whose words have helped me developed professionally: Prof. Kim Christensen, Dr. Tim Evans, Dr. Nick Jones, Dr. Steve Chum, my colleagues in DSO National Laboratories and my father.

Chuan Wen, Loe

Table of contents

List of Publications & Preprints

The following publications are based on the materials presented in this thesis.

[110] Chuan Wen, Loe and Henrik Jeldtoft Jensen
Bibliographic Search with Mark-and-Recapture
Physica A: Statistical Mechanics and its Applications, Accepted in 2015
doi: 10.1016/j.physa.2015.04.019

[111] Chuan Wen, Loe and Henrik Jeldtoft Jensen
Comparison of Communities Detection Algorithms for Multiplex
Physica A: Statistical Mechanics and its Applications, Accepted in 2015
doi: 10.1016/j.physa.2015.02.089

[114] Chuan Wen, Loe and Henrik Jeldtoft Jensen
Edge Union of Networks on the Same Vertex Set
Journal of Physics A: Mathematical and Theoretical, Vol. 46, No. 24, 2013.
doi:10.1088/1751-8113/46/24/245002

The following pre-prints are based on the materials presented in this thesis.

[112] Chuan Wen, Loe and Henrik Jeldtoft Jensen
Revisiting Interval Graphs for Network Science
Accepted by *Journal of Complex Networks*
doi: 10.1093/comnet/CNV023
arXiv:1503.07199

[113] Chuan Wen, Loe and Henrik Jeldtoft Jensen
Centrality of Unions of Networks on the Same Vertex Set
arXiv:1309.6629

Chapter 1

Introduction

Complexity Theory studies the collective behavior of a system of interacting agents, and a graph (network) is often an apt representation for such system. Typically the agents are modeled as indistinguishable vertices, and the interactions between a vertex pair are denoted with an edge in the graph. For example a social network maps people as vertices and their acquaintanceships as edges of a graph.

Although the theories and algorithms had shown to be successful in many applications, graphs occasionally trivialize the complex interactions between real-world entities. The sophisticated relationships between people are more than acquaintanceship, i.e. they can be colleagues, family, friends, etc.

Therefore it is important to refine the graphs such that we can encapsulate these relationships into our models. Broadly speaking this thesis studies the properties of a graph when additional edges are introduced, as *fine structures*, to represent these relationships.

Although such multi-relational systems are currently the modern outlook in Network Science, this thesis brings us back to the 1960s where Graph Theorists pondered about the same problems but with a different type of fine structures. We will study the nature of these fine structures and fill the gaps between Network Science and Graph Theory.

1.1 In Search for Linear Fine Structures of Life

The vertices of an interval graph represent intervals over a real line where overlapping intervals denote that their corresponding vertices are adjacent. This implies that the vertices are measurable by a metric and there exists a linear structure in the system.

Interval graphs was first applied to deduce the linearity of genes when Benzer noticed that the behavior of mutated strains of bacteriophage T4 (virus) forms an interval graph [15] *. The vertices are the different mutated variants of T4 such that they do not have the complete genome to kill bacterias independently. However two disabled viruses are able to so *together* if their mutated regions do not overlap, since the information of the entire genome of the original T4 is contained in the virus pair.

In the experiment, Benzer placed an edge between mutant pairs if the bacterias survived to denote that the pairs' mutated regions overlapped. The resultant graph was an interval graph which lead him to conclude the linear structure of genes. On the contrary if the resultant graph is not an interval graph, its topology can be determined in higher dimensions (example in section 2.1.3). This is used in ecology and operations research to deduce the "hidden" structures and stability of complex systems [50].

When these "hidden" structure are non-linear, it is the generalization of interval graphs where the vertices are d-dimensional hyper-boxes such that intersecting boxes imply that their corresponding vertices are adjacent in the graph G. This can be visualized by taking the species in an ecology as boxes and each of the axes measures a different environmental factor like temperature, soil acidity, amount of sunlight, etc. Each species are enclosed in their unique environment phase space where they are adaptable, and hence intersecting boxes imply that the species can coexist in a common environment.

Most of the research were published between the 1960s to 1980s where Graph Theorists study the non-linearity of the different graph ensembles (section 2.1). The reason is that to determine the dimensionality on general graphs is reducible to a NP-complete problem. Thus the research focus was shifted from a scientific framework to an analytical and computational interests of mathematicians.

---

*This thesis's title in fact pays homage to Benzer's paper, "On the topology of the genetic fine structure"

1.2 A Distant Family in Network Science

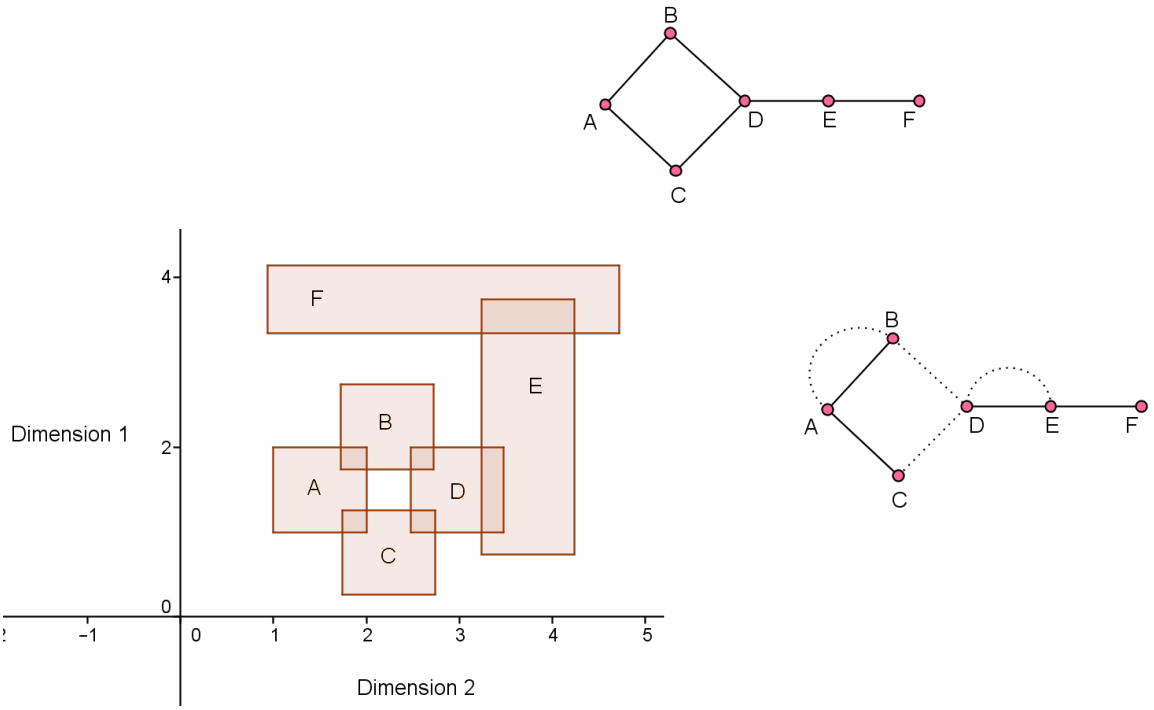

Figure 1.1: The bottom figures are the different fine structures of the graph in the top figure. The bottom left is the graph's 2-dimensional hyperbox representation, whereas the bottom right is the graph's multiplex on two relationships (represented by dotted and solid lines).

After interval graph research went out of fashion in the 1980s, it was not revisited when Network Science expressed interests with multi-relational[†] networks known as *multiplex*. A multiplex generalizes the edge set of a network where a vertex pair can be connected by multiple edges and each edge represents a different relationship. For example a social multiplex refines the relationships between people from acquaintanceship to specific roles like friends, colleagues or family (Fig. 1.1).

Multiplex research gain global populace in science as it is a natural transition from networks to preserve the rich relational data of the system [23, 95]. Hence multiplex research are very similar to network research, albeit more complicated analytically. For instance

[†]This is also known as multi-"dimensional" by some researchers. For clarity the term "dimension" is reserved for the hyper-boxes representations.

similar to networks, we can extend the study to the structural properties [17], information propagation [108], communities detection [111] and link prediction [56] of multiplexes.

Therefore there is no paradigm change from network to multiplex research, and hence there are very few reasons for one to refer back to Graph Theory (the origins of Network Science) for inspirations — many could have assumed that the change is equally small for Graph Theory. However this assumption is wrong as the dimensionality of graphs can be posed as a very different problem, i.e. interval graphs.

The connection is subtle as interval graphs and multiplexes are in fact a type of *intersection graph* [119], where vertex pairs are connected if they have overlapping attributes. For instance in a social multiplex, two individuals are connect as schoolmates if they are in the same school together (attribute); Similarly an interval graph of historical figures connects two people as contemporaries if they exist in the same period (attribute).

1.3 Outline

There is a clear gap in the relevant research of multiplexes and interval graphs (chapter 2). Thus chapter 3 detailed the structural properties found during the exploratory phase of this research and the connections between these fine structures.

Chapter 4 studies the communities detection problem on multiplexes. It is one of the main applications in Network Science to modularize a complex system into simpler components. There are multiple contributions in this chapter, namely a comprehensive survey and comparison of existing algorithms. New analytical bounds for the algorithms were also found using Probabilistic Methods [2] from Extremal Graph Theory.

Chapter 5 presents interval graphs from the perspective of Network Science. Using communities detection as a heuristic strategy, we are now able to determine the dimensionality of graphs more efficiently. Furthermore interval graph is also a plausible model to simulate the behavior of discontinuous flow of information in real world networks, i.e. information flow between nonadjacent vertices.

The compilation of such a large body of research is such a challenging task that Chapter 6 is dedicated to describe the process. The novelty of the chapter is the application of *Mark-and-Recapture* from population biology to estimate the minimum size of the essential

literature, i.e. measuring the completenesses of the bibliography.

Finally Chapter 7 summarizes this thesis and shares some personal perspectives. It highlights the challenges and uncertainties that arise during the investigations which in turn shapes the research trajectory of this thesis.

1.4 Research Trajectory and Contributions

The history of this research may help to explain the unfolding of this thesis's content and direction. Initially the focus was on the communities detection problem for multiplexes and chapter 4 was the phase to compare all the existing algorithms. However multiplex research is still in its infancy and there was not many open dataset or synthetic multiplex benchmarks for the study then.

Although new data can be collected, it is hard to support its quality and validity. Thus the decision was to investigate if established real-world networks (e.g. Zachary Karate Club Network) can be "reversed" such that we can derive the relationships for the multiplexes. This allows us to verify if the conclusions from the networks' algorithms are consistent (to a certain degree) with the results from the multiplexes' algorithms. I.e. since these networks are well studied, they act as the "ground-truth" for multiplexes.

Therefore chapters 3 and 5 are part of the research to understand the fundamentals of multiplexes with reference to networks. Unfortunately more research are still required before the goal will be met, however at this stage we had learned much more structural properties towards the transition from networks to its fine structures. Thus this thesis as its title suggested studies "the topology of network fine structures".

Finally we need to list down the main contributions in this thesis. The rational is that the contributions are disparate and easily lost within the text of this thesis. This is an interdisciplinary research where many somewhat unrelated ideas are interwoven together, thus it might not be clear to identify the new materials.

- Chapter 3 is mainly some analytical properties of the resultant graph when we project multiplexes on two relationships. All the materials are original except for Theorem 3.5.1, 3.5.2 and 3.6.1.

- Many of the results in Chapter 3 are exploratory findings, and hence only some are used in later chapters. I.e. section 3.4.1, 3.4.4 and 3.6 are foundations used in constructing some of our synthetic mathematical models in the later chapters.

- There are multiple contributions in Chapter 4:

  - The literature review on multiplex communities detection is more comprehensive and supplements the reviews by Boccaletti *et al.* and Kivelä *et al.*. [23, 95].

  - Derived new analytical bounds (corollary 4.3.2 and 4.3.3) for multiplex communities detection algorithms using probabilistic methods.

  - Showed that all the proposed multiplex communities detection algorithms are similar conceptually in ideal situations, but empirically very different when tested against benchmark multiplexes (Section 4.7).

- One of the challenges in creating benchmarks for multiplex communities detection algorithms is to determine the number of relationships in a multiplex. Many current literature worked around this issue by prudent decisions and qualitatively argue their choices. Section 5.1 shows that the connection between multiplexes and interval graphs allows us to quantify the number of relationships in a system (multiplex).

- Section 5.1 also introduced a heuristic method from Network Science to optimize the computation hard (graph theory) problem on interval graphs. More importantly this method is tolerant to experimental errors, hence meaningful for scientific work.

- Another main challenge in multiplex research is that the literature is disparate and thus it is difficult to consolidate the relevant research. Chapter 6 is an objective way to support the completeness of this thesis's bibliography using methodologies (i.e. Mark-and-Recapture) from population biology.

- Lastly the additional novelty in Chapter 6 is to apply Mark-and-Recapture on search engines to determine a stopping rule for research, i.e. how many entries in the search results must we explore before there is diminishing return in knowledge gained.

Chapter 2

Related Work

Interval graphs were actively studied during the 1960s and a comprehensive review can easily covers two book volumes [63, 119]. Similarly to consolidate the literature on multiplexes is also a massive task [23, 95], as the research is disparate across the different disciplines like Physics, Computer Science and Sociology. Hence to support the core materials of this thesis, only the relevant materials are presented in this review.

2.1 Interval Graphs

Definition 2.1.1 *An Interval Graph $I(V, E)$ maps a set of intervals $\{J^1, \ldots, J^n\}$ as vertices such that adjacent vertices $(a, b) \in E$ denotes $J^a \cap J^b \neq \emptyset$ [63] (Fig. 2.1).*

The sequential nature of the intervals implies that there is a linear ordering $\prec$ on the vertices where for all vertex triples $v_1, v_2, v_3 \in V$ with $v_1 \prec v_2$ and $v_2 \prec v_3$, if $(v_1, v_3) \in E$ then (v_1, v_2) or $(v_2, v_3) \in E$. This colloquially means that there is no "shortcut" in the graph, i.e. no independent vertex triples where every pair are connected by a path avoiding all neighbors of the third. This property is known as *Asteroid-Triple free* (AT-free).

Theorem 2.1.2 *An interval graph is chordal and AT-free [106].*

The lack of "shortcut" in AT-free graphs restricts the number of paths among the vertices in the graph and hence limits the search space for a variety of problems. Thus the AT-free property presents useful algorithmic structure on interval graphs such that some NP-complete graph problems are tractable in polynomial time [42].

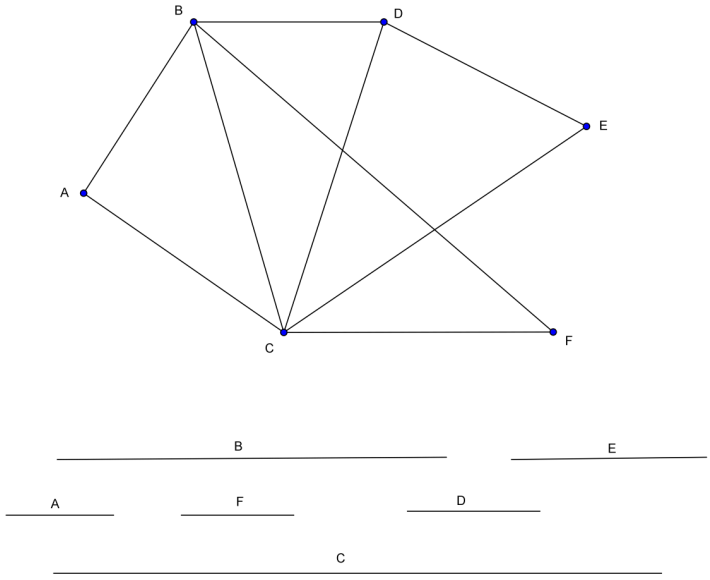

Figure 2.1: The duality of an interval graph (above) and a set of intervals (below). There is a bijective map between the vertices of the graph and the intervals where overlapping intervals denote the adjacency of their corresponding vertices. For example interval A overlaps interval B implies that vertex A is adjacent to vertex B, vice versa.

2.1.1 Recognizing Interval Graphs

A sketch of the algorithm to identify interval graphs helps to illustrate the intrinsic beauty of the AT-free property. It is similar to divide-and-conquer algorithms like quick-sort [81] where a pivot is chosen so that the problem is divided into more manageable search space.

In this case we need to choose a clique as a pivot such that it divides the rest of the vertices into two sets of intervals. Since there is no shortcut, the two sets of intervals are disjoint. This pivot is chosen from the list of maximal clique in the graph.

For all vertices v, the final step is to order the cliques in the list such that cliques with common v are arranged sequentially. If there is no such arrangement, then the graph is not an interval graph. Specifically a graph is an interval graph if there exists a column permutation on the incidence matrix between the cliques (columns) and vertices (rows) such that the ones in every row appears consecutively.

This can be computationally expensive without the use of PQ-Trees [26]. Hence consider the following simple example in Fig. 2.1: The list of maximal cliques of the graph is

$\{\{A, B, C\}, \{B, C, F\}, \{C, D, E\}, \{B, C, D\}\}$ and the incidence matrix is:

	ABC	BCF	CDE	BCD
A	1			
B	1	1		1
C	1	1	1	1
D			1	1
E			1	
F		1		

Row B does not have consecutive ones, as there is a gap in the column $\{C, D, E\}$. Hence by swapping the last two columns on the right, all the rows are in consecutive runs of ones and thus demonstrate that Fig. 2.1 is an interval graph.

2.1.2 Interval Graphs As Hyper-Boxes

The graph from the intersection of interval graphs $I^i(V, E_i)$, i.e. $G(V, E_1 \cap \ldots \cap E_m)$ forms a set of axis-parallel hyper-boxes as vertices in m dimensions, and adjacent vertices imply that their corresponding hyper-boxes intersect. The minimum m interval graphs to represent G is its *boxicity* and it is a measure of complexity (Fig. 2.2).

A *competition graph* in ecology connects two species (vertices) if they compete over the same food. For instance in Fig. 2.2, vertices A and B are connected, hence in the hyperbox representation (left of the graph) box A intersects box B. In the context of a marine food web, the axes can be described by two factors — the size of the prey and the depth of the water. I.e. the dynamics of the prey-and-predator can be modeled by the two niches, where each species is enclosed in its unique position of environment niches.

For example suppose the horizontal axis refers the size of the prey and the vertical axis refers to the depth of the water. Comparatively to species B, species A tends to be found nearer to the surface of the water and prey on smaller food sources. However since the boxes overlap, it means that at certain depth of the water we can find both species *and* food source of size that both of them can prey on. Although ecology is generally known to be a complex system, this hyperbox representation is conceptually much simpler. In fact Cohen showed that many food webs are interval graphs (1-dimensional) where the ordering of the intervals (as predators) correlates to the size of their preys [47].

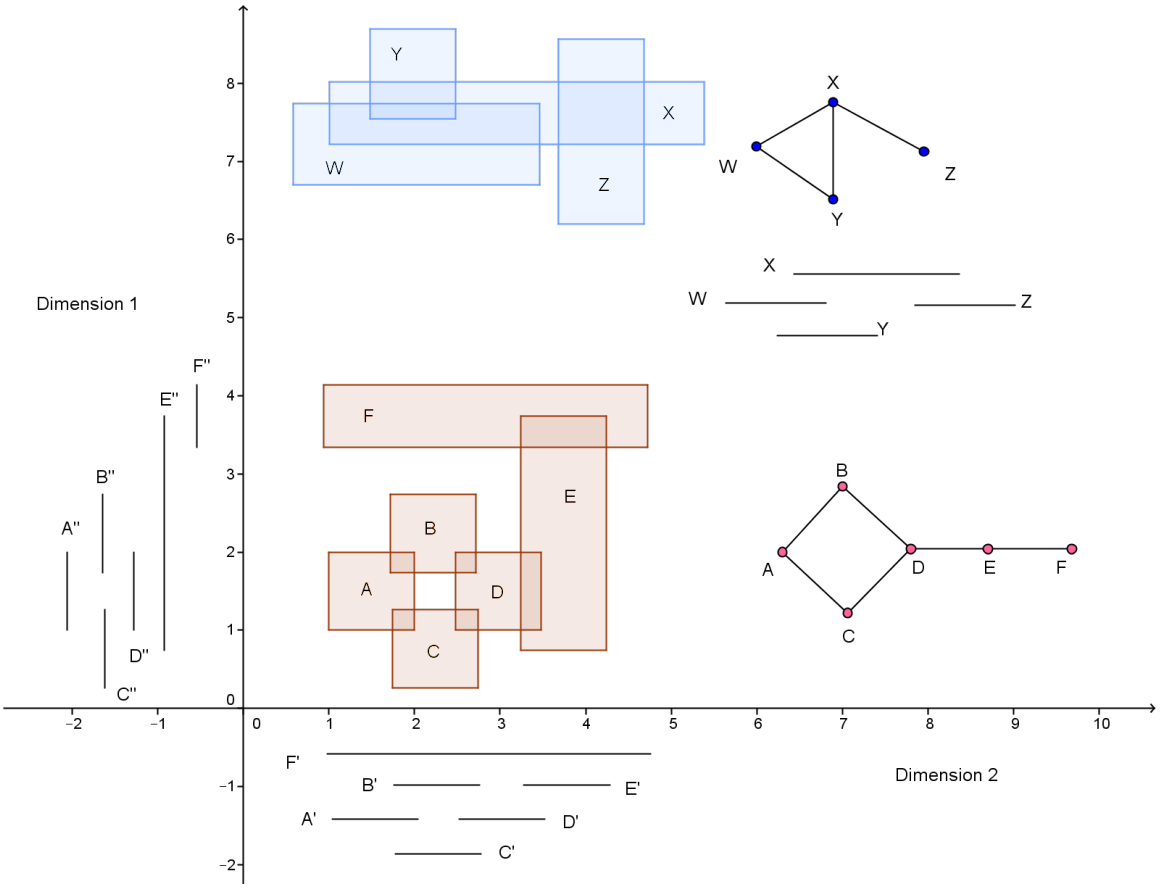

Figure 2.2: The set of 2-dimensional boxes $A, \ldots, F$ corresponds to the graph on its right, and they are from the intersection of 2 interval graphs with vertex labels $A', \ldots, F'$ (dimensional 1) and $A'', \ldots, F''$ (dimensional 2). Adjacent vertices are equivalent to saying that their respective boxes intersect. However a graph constructed from m-dimensional boxes does not necessary implies that it has boxicity m. For instance the top graph with vertex labels $X, \ldots, Z$ is constructed with $m = 2$-dimensional boxes, but since it can be represented with a 1-dimensional interval graph, its boxicity is one.

However it is computationally hard to determine the boxicity of an arbitrary graph [140] as there are more degrees of freedom to position the hyperboxes in d-space. For example to test if the bottom graph in Fig. 2.2 can be embedded in 2 dimensions, the exhaustive search algorithm is to iterate all pairs of interval graphs in hope that there is a pair such that their (edge) intersection gives us the graph. Each possible candidate interval graph on n vertices is first determined by the ordering of n disjoint intervals (e.g. $A'B'C'D'E'F'$). Next the intervals are extended such that the interval graph's edge set is the superset of the graph's edge set. Since there are $n!$ possible orderings in the first step, hence there are $(n!)^2$ pairs of candidate interval graphs solutions. If no solution is found, then the process is repeated for all 3 (more if necessary) combinatorial tuples of interval graphs to check if the graph can be embedded in 3 (or more) dimensions. Hence it was an active research to bound the boxicity of graphs (Table 2.1) such that computation can be bounded:

Graph	Boxicity
Cycle [140]	$= 2$
Tree [41]	$= 2$
Outerplanar graph [149]	≤ 2
Planar graph [162]	≤ 3
Bipartite graph with independent sets V_1 and V_2 [41]	$\leq min\lceil\frac{\lvert V_1\rvert}{2}\rceil, \lceil\frac{\lvert V_2\rvert}{2}\rceil$
Graph with minimum vertex cover of size t [41]	$\leq \lfloor\frac{t}{2}+1\rfloor$
Turan graph on n vertices with $n/2$ partitions [41]	= $n/2$
Split graph with clique K [41, 51]	$\leq \lceil\frac{\lvert K\rvert}{2}\rceil$
Complete multipartite graph $K_{n_1,\ldots,n_p}$ [140]	$= \lvert i : n_i > 1\rvert$
Graph with genus g [62]	$\leq 5g+3$
Line graph of a multigraph with maximum degree d [43]	$\leq 2d(\lceil log_2(log_2 d)\rceil + 3) + 1$
Graph on n vertices with average degree d [42]	= $O(d \ln n)$
Graph on n vertices with Maximum degree d [44]	$\leq min(n/2, d^2+2, \lceil(d+2)\ln n\rceil)$
Graph on n vertices with Minimum degree d [1]	$\geq n/(2(n-d-1))$

Table 2.1: Boxicity of different graphs ensembles.

However most of these bounds are not tight, e.g. a balance bipartite graph has boxicity up to one quarter of the graph's size. Amongst the list of analytical bounds, there is an interesting relationship between boxicity and graph genus. The genus of a graph is defined as the minimum number of holes on a surface such that the graph can be embedded without crossing edges (planar). What is noteworthy is that the genus of a graph affects the scaling

properties from large-world (small genus) to ultrasmall-world (large genus) networks [6]. Consequently a graph with large boxicity implies an ultrasmall-world network, although conversely it is not true as low boxicity does not imply large-world network (e.g. a complete graph has boxicity 1, but with genus $= \lceil (n-3)(n-4)/12 \rceil$) [6].

Lastly the boxes are synonymous to embedding graphs in m dimensional Minkowski r-metric space M_m^r such that for all adjacent $u, v \in V$, their distance in the metric space is bounded by some length [66]:

$$d_{uv}(\langle f_1(u), \ldots, f_m(u) \rangle, \langle f_1(v), \ldots, f_m(v) \rangle) \leq l_u + l_v, \tag{2.1}$$

where l_u and l_v are length given for their respective vertices*, and $\langle f_1(u), \ldots, f_m(u) \rangle$ is a vector mapping u to the metric space with the real-value functions $f_1, \ldots, f_m$. In addition the functions f_i on all $u, v \in V$ is conditioned by Minkowski r-metric space:

$$d_{uv} = \Big[\sum_{i=1}^{m} |f_i(u) - f_i(v)|^r \Big]^{1/r}. \tag{2.2}$$

The arbitrary constant r is a weighting parameter where all components $|f_i(u) - f_i(v)|$ are equally weighted for $r = 1$ (i.e. Manhattan Distance). For $r = 2$ (i.e. Euclidean Distance), the components that are greater contribute more to the distance. Hence by letting $r = \infty$ to complete the metric space, the greatest component will dominate i.e. $d_{uv} = \max_{i=1}^{m} |f_i(u) - f_i(v)|$, where each vertex is a hyper-box with sides parallel to the axes.

2.1.3 Topology of an Unknown Structure (Example)

How did Benzer deduced the linearity of genes with interval graphs (backstory in section 1.1)? And how is it different from a non-linear structure? Suppose there are two hypotheses of a gene's structure — linear and branched (Fig. 2.3).

*Each vertex is given a distinct length that corresponds to its "volume" of the hyperbox. Suppose for all u and v, their length are equal, i.e. $l_u = l_v = l$, then in $r = 2$ each vertex is a sphere with radius l. Vertex pairs are adjacent if their corresponding spheres intersect. In the case of boxicity, the volume hyperboxes varies in volume, hence l_u do not necessarily equals to l_v. Hence l_u and l_v are arbitrarily chosen such that they are long enough to intersect with adjacent vertices' hyperboxes, yet short enough to avoid intersecting non-adjacent vertices' hyperboxes.

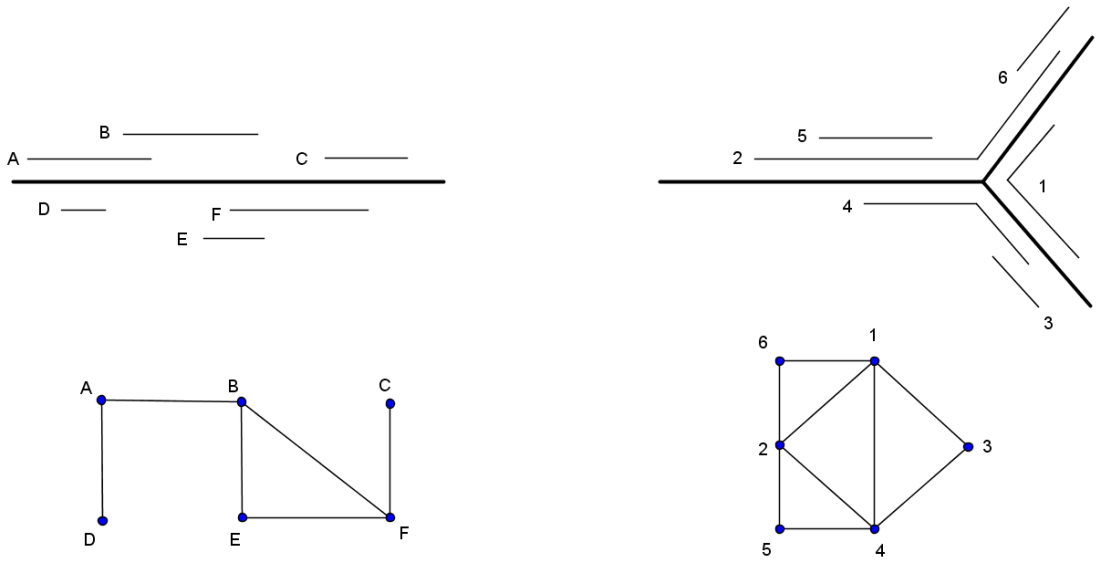

Figure 2.3: A comparison of a linear structure (left) and a branched structure (right). Adjacent vertices denote that their respective segments overlap (e.g. vertex C is adjacent with vertex F as segment C overlaps with segment F). Since the graph on the left corresponds to a linear structure, it is an interval graph. The graph on the right is not an interval graph as vertices 3, 5 and 6 form an asteroid-triple — path 3-1-6 (3-4-5 and 5-2-6) avoids the neighbors of vertex 5 (respectively 6 and 3).

The vertices are the different mutated variants of a virus such that they do not have the complete genome to kill bacterias independently. Let vertex A refers to the variant where segment A of the virus is changed. Given that virus pairs with overlapping segments do not have the complete information to kill the bacterias, an edge is placed between them. Since the graph on the left is constructed from a linear structure, we get an interval graph.

However if genes were a branched structure, then the resultant graph will not be an interval graph. In the same figure, vertices 3, 5 and 6 form an asteroid-triple — the path 3-1-6 (3-4-5 and 5-2-6) avoids the neighbors of vertex 5 (respectively 6 and 3). It is noteworthy to observe that by removing any of the vertices 1, 2 or 4 is sufficient to reduce the graph to an interval graph. It is also possible to get an interval graph by removing edges $\{1, 3\}$ and $\{1, 4\}$ from the original graph. Thus interval graphs only *supports* the hypothesis of a linear structure, but it is insufficient to *prove* the linearity of a system.

Therefore interval graphs is a very sensitive to changes and experimental errors. Section 5.1.3 shows that communities detection can be used to identify graphs that can be expressed as interval graphs with minor modifications.

2.1.4 Scale-Free Interval Graph

Miyoshi *et al.* proposed an interval graph ensemble that exhibits two real-world characteristics — high clustering coefficient and power-law degree distribution. However due to the Asteroid-Triple-Free property of interval graphs, it fails to be a small-world graph [121].

The construction (Fig. 2.4) is similar to a Barabási-Albert graph (section 2.3.1) where the scale-free graph grows by iteratively adding new vertices to the graph. The number of new vertices at each time step follows a Poisson distribution with mean λ and their length L follows a power-law distribution $Pr(L = k) = 1/\zeta(\alpha) \cdot (k + 1)^{-\alpha}$, where $\zeta(\alpha)$ is the Riemann's zeta function. Hence the probability that a vertex has degree k follows a power-law distribution $P(k) \sim \frac{\lambda^{\alpha-1}}{\zeta(\alpha)} k^{-\alpha}$, and the clustering coefficient is bounded below by

$$Pr(L = 0) + Pr(L = 1)e^{-C_\lambda}\Big(1 + C_\lambda + \frac{1}{2}\sum_{d\geq 2} \frac{d-2}{d-1}\frac{C_\lambda^k}{d!}\Big), \tag{2.3}$$

where $C_\lambda = \lambda\Big(\frac{\zeta(\alpha-1)}{\zeta(\alpha)} + \Big) - 1$ [121].

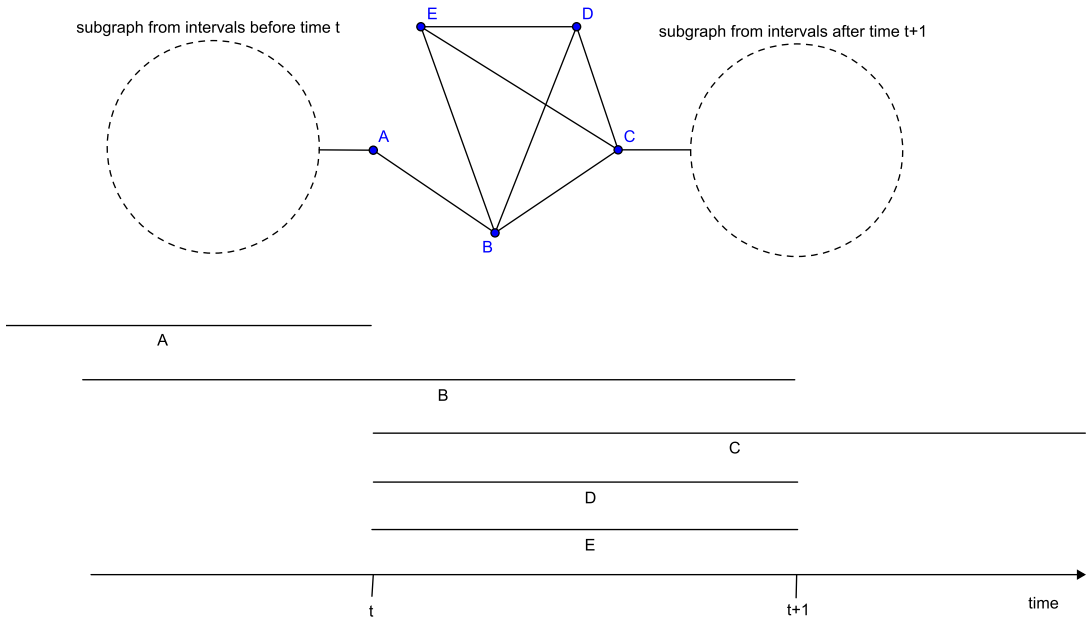

Figure 2.4: Visual description on the construction of scale-free interval graph with high clustering coefficient. The underlying linear structure is time (axis at the bottom). It is similar to Barabási-Albert graph where at time t, we add some new vertices/intervals. The number of new vertices is determined by the Poisson distribution where in this example 3 new intervals are added (interval C, D and E). Next, the length of the new intervals are determined by a power-law distribution and in this example length of interval D and E is 1 unit of time. The above is the interval graph from the set of intervals. In the complete implementation in [121], there is an addition procedure that extends the length of intervals generated before time t. At time t, intervals with end points at time t, e.g. interval A, can be extended to time $t + 1$ *probabilistically* (determine by power-law distribution). In this example, only interval B is extended. This procedure allows more intervals to have longer length and hence more likely for the graph to be connected. Without this additional procedure, most intervals will have unit length (since the length is power-law distributed) and will have many instances like disjoint intervals A and D.

2.1.5 Real World Examples

Besides in Bioinformatics [20, 87], interval graphs arise naturally in many time dependent applications like task scheduling [39] or other linear structures like pavement deterioration analysis [68] and food-webs in ecology [47].

Scheduling is one of the main problems in operation research where it optimizes the process to complete a set of tasks, where every task is an interval with a specific start that cannot be interrupted. Thus it can be modeled as an interval graph with the tasks as vertices and two vertices are adjacent if their corresponding tasks overlap in time. For example if each task can only be assigned to one machine, the *basic interval scheduling problem* questions the minimum number of machines required to complete all the tasks. This can be posed as a graph theory problem to determine the chromatic number of the interval graph, where each color defines a distinct machine [96].

In ecology, a system is stable if few species will go extinct after random species are removed from the ecology. It is believed that low complexity food-webs are generally less stable [86], and boxicity is one of such measures [47, 158].

For example a food-web that is an interval graph usually orders the species according to the size of their prey. If all the species of a particular size is removed from the ecology, their predators which are usually slightly larger in size will in turn dies off. Since the effects from the extinction of these predators will cascade to larger predators, thus causing the instability of the ecology.

Therefore the boxicity of an ecology is of particular interests for ecologists. Eklöf *et al.* analyzed 200 ecological networks and discovered that only 35 systems are interval graphs, and almost all but four have boxicity < 6 [60]. The maximum boxicity is the Phrygana Pollination network with boxicity $= 9$. It was also observed that the boxicity scales almost linearly with the logarithm of the number of edges.

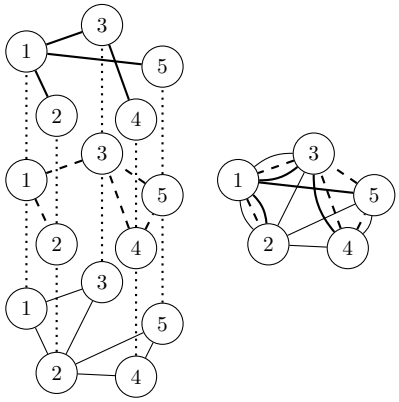

Figure 2.5: The two different visual representations of the same multiplex $\mathcal{M} = \{G^1, G^2, G^3\}$. The left figure shows 3 "layers" of graphs where G^1, G^2 and G^3 are the upper, mid and lower layer respectively. Each relationships is drawn with a distinct line style. The figure on the right is a "flatten" representation of the same multiplex, where parallel edges are drawn between vertex pairs to illustrate the multiple relationships. The dotted vertical lines are just a way to visualize the alignment of the vertices in the layered representation of the multiplex.

2.2 Multiplex

2.2.1 Disparate Terminology

Definition 2.2.1 ***(Multiplex)*** *A multiplex is a finite set of m graphs, $\mathcal{M} = \{G^1, \ldots, G^m\}$, where every graph $G^i = (V, E_i)$ has a distinct edge set $E_i \subseteq V \times V$ (Fig.2.5).*

Multiplex is a natural transition from graph for researchers as a model to preserve the rich relational properties in the data. Hence there are concurrent and disparate investigations on multiplex which inevitably gave multiplex many synonymous names like: **Multigraph** [49, 71], **MultiDimensional Network**[16, 17, 83, 92, 107, 146, 161], **Multi-Relational Network** [33, 36, 115, 143, 153, 159], **MultiLayer Network** [28, 31, 52, 108, 123], **PolySocial Networks** [5], **Multi-Modal Network** [3, 103, 126], **Heterogeneous Networks** [56] and **Multiple Networks** [117]. The fragmented nature of the literature will be further discussed in Chapter 6.

2.2.2 Projection Of Multiplex

Definition 2.2.2 ***(Projection)*** *Given a multiplex $\mathcal{M} = \{G^1(V, E_1), \ldots, G^m(V, E_m)\}$, its projection is the graph $G(V, E_1 \cup \ldots \cup E_m)$.*

One of the strategies to understand high dimensional systems is to map the system onto lower dimensions. The projection of a multiplex is a simpler and more familiar representation to work on. For example Zachary Karate Club Network is a projection of eight social relationships among the karate club members [167].

Although we can get more "sophisticated" by proposing similar graph metrics for multiplex, it only creates further fragmentation to the literature. For example the following are alternative extensions for a multiplex to define the set of *adjacent* vertices of vertex v, i.e. the neighbors of v:

Definition 2.2.3 ***(OR-Neighbors)*** *Let $M^* \subseteq \mathcal{M}$, the OR-neighbors of $v \in V$ is the number of vertices adjacent to v in at least one of the graph in M^* [17, 146]:*

$$\textit{OR-Neighbors}(v, M^*) = \{u | (u, v) \in E_i, G^i \in M^*, u \in V\}. \tag{2.4}$$

Definition 2.2.4 ***(AND-Neighbors)*** *Let $M^* \subseteq \mathcal{M}$, the AND-Neighbors of $v \in V$ is the number of adjacent vertices to v in all $G^i \in M^*$ [19]:*

$$\textit{AND-Neighbors}(v, M^*) = \{u | (u, v) \in \cap_{G^i \in M^*} E_i, u \in V\}. \tag{2.5}$$

Definition 2.2.5 ***(XOR-Neighbors)*** *Let $M^* \subseteq \mathcal{M}$, the XOR-Neighbors of $v \in V$ is the number of vertices adjacent to v in M^* and not in the complement of M^* [17]:*

$$\begin{aligned} &\textit{XOR-neighbors}(v, M^*) = \\ &\quad \{u | (u, v) \in \cup_{G^i \in M^*} E_i \setminus \cup_{G^j \in \mathcal{M} \setminus M^*} E_j, u \in V\}. \end{aligned} \tag{2.6}$$

Definition 2.2.6 ***(MIN-Neighbors)*** *The MIN-Neighbors of $v \in V$ is the number of vertices $u \in V$ that are adjacent to v in at least α graphs [30]:*

$$\textit{MIN-Neighbors}(v, \alpha) = \Big\{ u \Big| \sum_{G^i \in \mathcal{M}} \delta(u, v, E_i) \geq \alpha \Big\}, \tag{2.7}$$

where $\delta(u, v, E_i) = 1$ if $(u, v) \in E_i$, zero otherwise.

For example consider v to be the 3^{rd} vertex in Fig. 2.5 and let $M^* = \{G^1, G^2\}$. Thus the OR-Neighbors, AND-Neighbors and XOR-Neighbors of v are $\{1, 4, 5\}$, $\{1, 4\}$ and $\{4\}$ respectively. Also if $\alpha = |M^*| = 2$, then the MIN-Neighbor of v is $\{1,4\}$.

Each of these definitions can be used as the building blocks to algorithmically extend new metrics for multiplex [30, 31, 33, 107]. For instance the clustering coefficient of a vertex v is the ratio of the number of triangles at v to the number of pairs of adjacent vertices to v. Since a triangle is a cycle of 3 adjacent steps, the clustering coefficient of a multiplex can be extended to any of the definitions of *adjacency* in multiplex [30, 53].

Although these new definitions might be meaningful to their respective applications, they are too narrow to study multiplex purely as a general mathematical object, yet too diverse if every combinations of the definitions are considered in this thesis. Thus this thesis does not further investigate these alternative definitions.

2.2.3 Multiplex Communities

A *community* is broadly described as a set of interacting agents that collectively behaves differently with non-community agents. The process to identify these communities modularizes a complex system into simpler representations so as to form the bigger picture of the system. However there is no universally accepted formal definition as the construct of a community often depends on its problem domain [64]. The relevant literature on multiplex communities are presented in chapter 4 so that the chapter can be self-contained.

2.2.4 Real World Examples

Depending on the level of abstraction, many applications can be disguised as multiplexes [23, 95] or other generalized graph models (section 6.1). Thus to avoid introducing unnecessary technical terminologies, it will be more accessible to take commonplace examples like social networks and transportation systems.

One of the early references of multiplexes in social networks was to describe the "multiplexity" social ties of the Medici family in the 1400s [133]. However only recently there are scientific research on multiplexes to study the dynamics of people that may be connected through more than one form of relationships, e.g. family, hobbies, work place and so forth. Since it is a closer representation of the real world system, it is assumed that it will be more accurate to determine the community structures [16, 18, 58, 105, 120, 134, 161] or to predict future links between people [76, 109].

However the type and the number of relationships in a social multiplex is often arbitrarily chosen by the researchers or limited by the quality of the data. For example the Zachary Karate Club Network [167] was from the data of eight different types of relationships and it was projected as a single-relational network. The granularity of these relationships is important as it has effects on simple features like the centrality [13], yet there are very few studies on it [55]. In fact many open dataset for social multiplexes [59, 141, 161] do not critically justify their choice of relationships.

In contrast the relationships in transport multiplexes are often well defined by the physical infrastructure. Multiplexes in transportation networks are also known as multimodal networks, where bus stops, train stations and terminals are indistinguishable locations (vertices) to transit to a different mode of travel. Cardillo *et al.* modeled the data of European Air Transportation (EAT) Network as a multiplex of airports where two of them are connected if there is a direct flight between them [37, 38]. To define the multiplexity of the data, each relationship maps the routes of a different airline.

2.3 Preliminaries

The following are some essential basics that are mentioned throughout this thesis.

2.3.1 Graph Ensembles

Erdős-Rényi, Watts-Strogatz and Barabási-Albert models are some of the popular graph ensembles in Network Science for their simple constructions and useful mathematical properties. This thesis used different combinations of these graph ensembles as constructions for multiplex.

Erdős-Rényi

A realization of an Erdős-Rényi graph [61] is selected with equal probability from the set of all possible graphs with n vertices and $|E|$ edges. However to generate a huge random Erdős-Rényi is difficult. To circumvent this problem one may instead let the number of edges fluctuate slightly and consider a $G_{n,p}$ model in which every vertex pair is connected with probability p where $p = |E|/\binom{n}{2}$. The difference is that Erdős-Rényi has precisely $|E|$ edges while $G_{n,p}$ has approximately $|E|$ edges with high probability.

The properties on the edges and vertices of a $G_{n,p}$ graph can be easily expressed as random variables of well known distributions, hence there are much more analytical results on $G_{n,p}$ than Erdős-Rényi graph. For example if $lim_{n\to\infty} np \to constant$, then the degree distribution of the graph follows a Poisson distribution with mean np. However the most famous result among them is in the 1960 paper by Erdős and Rényi:

Theorem 2.3.1 ***(Sharp Threshold of*** $G_{n,p}$***)*** *Let $G_{n,p}$ be a graph on n vertices where vertex pairs are connected with probability p. The sharp threshold for connectedness is $\ln n/n$ such that If $p < \ln n/n$ (respectively $p > \ln n/n$), then with high probability $G_{n,p}$ is disconnected (respectively connected) [61].*

Therefore many literature also refer Erdős-Rényi graph as $G_{n,p}$, although it was proposed much earlier by Solomonoff, Rapaport and Gilbert in 1950s [69, 154].

Watts-Strogatz

A Watts-Strogatz graph on n vertices, $W_{n,w,q}$ is parameterized by w and q for its mean degree and probability of rewiring respectively [164]. The construction begins with a regular ring lattice where each vertex connects to $w/2$ neighbors on each side.

Let the vertices be ordered from $v_1, \ldots, v_n$. For every vertex $v_i \in V$ and $i < a$, each edge leaving v_i is rewired with probability q. The rewiring process replaces $\{v_i, v_a\}$ with $\{v_i, v_b\}$ where v_b is chosen uniformly in the set $\{v_{a+1}, \ldots, v_n\}$ such that the resultant graph remains a simple graph and every configuration has an equal chance to occur. This process allows a Watts-Strogatz graph to evolve from a regular Ring Lattice to an Erdős-Rényi graph, i.e. as $q \to 1$, $W_{n,w,q} \to G_{n,w/(n-1)}$.

The distinctive property of a Watts-Strogatz graph is its high clustering coefficient that decreases at the rate $(1-q)^3$ for increasing value q. For $q = 0$, its clustering coefficient is $\approx 3/4$ and it is independent to the size of the graph n [11].

Barabási-Albert

Barabási-Albert graph on n vertices [10] is denoted by $B_{n,s}$ where s is the number of new edges at each iteration. The construction begins with some arbitrary small number of vertices connected at random.

At each iteration, one new vertex of degree s is added. The edges of the new vertex are connected probabilistically with a probability p_i proportional to the degree of the existing vertices v_i. This is referred to as the preferential attachment and is defined by:

$$p_i = \frac{degree(v_i)}{\sum_j degree(v_j)}. \tag{2.8}$$

This process yields a graph with power-law degree distribution with probability distribution function $\sim 2s^2k^{-3}$.

2.3.2 Louvain Algorithm

Definition 2.3.2 ***(Modularity)*** *Let A_{ij} be the adjacency matrix of a graph with $|E|$ edges and k_i is the degree of vertex i. Indicator function $\delta(v_i, v_j) = 1$ if v_i and v_j are in the same community, otherwise $\delta(v_i, v_j) = 0$. The quality (modularity) of the communities is defined by [130]:*

$$Q = \frac{1}{2|E|} \sum_{ij} \Big(A_{ij} - \frac{k_i k_j}{2|E|} \Big) \delta(v_i, v_j). \tag{2.9}$$

The *modularity* of a partition of a graph measures how different the clusters of vertices (communities) are from a random graph (Def. 2.3.2), and this global metric is simply the summation of local information between all vertex pairs which will face resolution limits [100]. The effectiveness of modularity function is limited by the size of the graph as local information becomes less representative to the bigger picture by ignoring larger but more meaningful subgraph structures. For example small and dense cliques that are loosely connected will be grouped together as a community by the modularity function although contextually it is preferred to identify these cliques as individual communities. Thus we will fail to discover well-defined small communities in large graphs. This is not a main concern in this thesis as the graphs in our experiments are usually small, within the order of thousands of vertices.

Louvain Algorithm [22] is one of the many communities detection algorithms in the literature to maximize the modularity function. It is a greedy algorithm that iterates a two-steps process that first optimizes the modularity locally into communities and then merges the vertices in the same community as vertices of a new (higher "hierarchy") graph. Each iteration is a hierarchy of partition and the hierarchy with the maximum modularity is chosen as the solution.

In the communities detection problem for multiplex (chapter 4), some algorithms require a communities detection algorithm for graphs as an interim step to derive a partial solution. Although it is assumed that any arbitrary algorithm is suffice, in some of the literature Louvain Algorithm was chosen [18, 166]. Hence to stay close to the experiments in the literature, the other communities detection algorithms for graphs [64] were not considered in this thesis.

Chapter 3

Statistical and Structural Properties of Multiplex and Interval Graph

The research presented in this chapter is published in [112, 113, 114].

3.1 Notations and Preliminaries

The following are the notations for the different graph ensembles in this thesis:

Notation	Description
$G(V, E)$	A graph on vertex set V and edge set E.
$\mathcal{M}$	A multiplex.
G^i	The i^{th} graph in the multiplex.
I^i	The i^{th} interval graph.
J^i	The i^{th} evolutionary interval graph.
$G_{n,p}$	Erdős-Rényi Graph.
$B_{n,m}$	Barabási-Albert graph.
$W_{n,w,q}$	Watts-Strogatz graph.
$R_{n,d}$	A d-regular graph.
$G_{n,\alpha}$	A random graph with maximum degree α.
G^P	The resultant graph from the projection of $\mathcal{M}$.
H	The resultant graph from the intersection of the edge set of $\mathcal{M}$.

In addition the following are the notations used exclusive for this chapter:

Notation	Description
$\cup$	The union of the edge sets (projection) of a multiplex.
$\cap$	The intersection of the edge sets of a multiplex.
v^i_{max}	The vertex with the greatest centrality in G^i.
v^P_{max}	The vertex with the greatest centrality in G^P.
$deg_i(v)$	Degree of vertex v in graph G^i.
$deg_P(v)$	Degree of vertex v in projection graph.
$T(v)$	The number of triangles incident to vertex v.

To ensure that none of the graphs in the multiplex dominate the behavior of the multiplex, the size of the edge sets has to be approximately equal, i.e. $|E_i| \approx |E_j|$. Lastly the following lemmas and concepts are useful in the later analysis.

Definition 3.1.1 ***(Hypergeometric Function)*** *Suppose we have n urns. Choose $a < n$ urns and place one ball in each. Next choose $b < n$ urns and place one ball in each. The probability that exactly i urns contain two balls is defined by the hypergeometric function*

$$\mathcal{H}(i; n, a, b) = \frac{\binom{a}{i}\binom{n-a}{b-i}}{\binom{n}{b}}. \tag{3.1}$$

Definition 3.1.2 ***(Overlaps)*** *Let two edges from two different graphs in $\mathcal{M}$ be $e(u, v) \in E_i$ and $e'(u', v') \in E_j$, where $i \neq j$. Edges e and e' overlap if and only if $e = e'$.*

Lemma 3.1.3 *Let E_1 and E_2 be the edge sets of two $G_{n,p}$. The probability that there are ϵ overlapping edges is:*

$$P(|E_1 \cap E_2| = \epsilon) = \mathcal{H}\Big(\epsilon; \binom{n}{2}, |E_1|, |E_2|\Big). \tag{3.2}$$

Proof Color $|E_1|$ edges blue out of the total possible $\binom{n}{2}$ edges of a complete graph to denote the edges of E_1. Next color $|E_2|$ edges red to the same complete graph. The probability that ϵ blue edges are colored over by red is defined by the hypergeometric function.

Eq. 3.2 is the probability that there are ϵ overlapping edges between the two graphs on the same vertex set. Thus the expected number of overlapping edges is given by:

Corollary 3.1.4 *Let E_1 and E_2 be the edge sets of two $G_{n,p}$. The expected number of overlapping edges is the expectancy of the hypergeometric function in equation 3.2:*

$$\mathbb{E}[|E_1 \cap E_2|] = \mathbb{E}[\mathcal{H}] = \frac{|E_1| \cdot |E_2|}{\binom{n}{2}}. \tag{3.3}$$

In fact we can generalize the above to compute the number of overlapping cliques in the union. In particular it is useful to count the number of overlapping triangles (3-clique) as it affects the accuracy of the clustering coefficient in the projection of two graphs.

Lemma 3.1.5 *Let K_1 and K_2 be the sets of c-cliques of two $G_{n,p}$. The probability that there are ϵ overlapping c-cliques is:*

$$P(|K_1 \cap K_2| = \epsilon) = \mathcal{H}\Big(\epsilon; \binom{n}{c}, |K_1|, |K_2|\Big). \tag{3.4}$$

Proof Color $|K_1|$ c-cliques clue out of the total possible $\binom{n}{c}$ c-cliques of a complete graph. Next choose $|K_2|$ c-cliques from the complete graph and color the edges red. The probability that ϵ blue c-cliques is colored over by red is defined by the hypergeometric function.

Corollary 3.1.6 *Let K_1 and K_2 be the sets of c-cliques of two $G_{n,p}$. The expected number of overlapping c-cliques is the expectancy of the hypergeometric function in equation 3.4:*

$$\mathbb{E}[|K_1 \cap K_2|] = \mathbb{E}[\mathcal{H}] = \frac{|K_1| \cdot |K_2|}{\binom{n}{c}}. \tag{3.5}$$

The above lemmas are some global measures on the projection of two Erdős-Rényi graphs and can be inaccurate for special ensembles of graphs. For example in the projection of two star graphs, lemma 3.1.3 is not applicable to estimate the distribution of the overlapping edges, since there are either $n-1$ or 2 overlapping edges.

However to compute the degree distribution of the projection graph accurately, we have to minimize the double counting of the overlapping edges. Although the global measures might not be accurate for the projection of non-Erdős-Rényi graphs, the same methodology can be applied at a local level such that the errors can be minimized.

Lemma 3.1.7 *Let $v_{i,1}$ and $v_{i,2}$ be the vertices of graphs G^1 and G^2 on n vertices respectively. If $d_1 = deg(v_{i,G^1})$ and $d_2 = deg(v_{i,G^2})$, then the probability that there are ϵ overlapping edges between $v_{i,1}$ and $v_{i,2}$ is:*

$$P_o(\epsilon|d_1, d_2) = \mathcal{H}(\epsilon; n-1, d_1, d_2). \tag{3.6}$$

Proof There are $n-1$ vertices left for $v_{i,1}$ and $v_{i,2}$ to connect to. Similar to the argument in Lemma 3.1.3, there are d_1 blue edges and d_2 red edges from $v_{i,1}$ and $v_{i,2}$ respectively.

Corollary 3.1.8 *Let $v_{i,1}$ and $v_{i,2}$ be the vertices of graphs G^1 and G^2 on n vertices respectively. Let $d_1 = deg(v_{i,1})$ and $d_2 = deg(v_{i,2})$, the expected number of overlapping edges between $v_{i,1}$ and $v_{i,2}$ is:*

$$\mathbb{E}[\mathcal{H}] = \frac{|d_1| \cdot |d_2|}{n-1}. \tag{3.7}$$

3.2 The Intersection of Multiplex

The projection of a multiplex appears to be the counter-thesis of network fine structures by reducing the problem back to a graph. However to understand the connection between multiplexes and graphs, it is important to study the process from both directions. Hence without compromising too much relational information for simplicity, the analysis of the overlapping edges pivotal, i.e. the graph $H(V, E_1 \cap \ldots \cap E_m)$.

The distribution of the overlapping edges is an essential characteristic to distinguish multiplex ensembles from random [19, 45, 104, 139]. For example a multiplex is correlated if the expected number of overlapping edges deviates from the projection of random Erdős-Rényi graphs (lemma 3.1.3) [19]. The degree of correlation in turn affects the phase transition of its structural properties like the emergence of a giant component [104].

Since each graph in a multiplex is the intersection of interval graphs, i.e. $G^i(V, E_i) \in \mathcal{M} = I_1^i \cap \ldots \cap I_d^i$, thus the set of overlapping edges is also the hyper-box representation of the system. Specifically the set of overlapping edges form the graph $H(V, E_1 \cap \ldots \cap E_m) = (I_1^1 \cap \ldots \cap I_d^1) \cap \ldots \cap (I_1^m \cap \ldots \cap I_{d'}^m)$. Hence the boxicity of H is not more than the sum of the boxicity of the graphs, i.e. boxicity$(H) \leq d + \ldots + d'$.

3.3 Degree Distribution

The degree (or valency) of a vertex is a measure of its connectivity where the high degree vertices are usually the important agents in a system, e.g. major airports. However a vertex's degree is insufficient to determine its relative value in a huge graph as it is a local measure. Thus the degree distribution of a graph is more informative.

The degree distribution $Pr(k)$ of a graph is the fraction of vertices with degree k. Therefore it is easy to determine a vertex's percentile ranking of its connectivity from the degree distribution. In addition the degree distribution is a global indicator to measure how similar a graph is to a real world system [10].

3.3.1 Erdős-Rényi with Erdős-Rényi

The null model of a multiplex is the set of Erdős-Rényi graphs, i.e. $\mathcal{M} = \{G_{n,p}, \ldots, G_{n,p'}\}$. Since the probability that an edge exists is independent for all graphs in the multiplex, hence the probability that an edge exists in the multiplex's projection of intersection is simply the product of probabilities. Hence the resultant graphs are also Erdős-Rényi graphs:

$$G^P = G_{n,p} \cup \ldots \cup G_{n,p'} \sim G_{n,1-(1-p)\ldots(1-p')}, \tag{3.8}$$

and

$$H = G_{n,p} \cap \ldots \cap G_{n,p'} \sim G_{n,p\ldots p'}. \tag{3.9}$$

Thus their degree distributions follow the Erdős-Rényi graph ensemble.

3.3.2 Erdős-Rényi with Watts-Strogatz

As $q \to 1$, a Watts-Strogatz $W_{n,w,q}$ can be approximated as an Erdős-Rényi [164], so in this limit the union is described by Eq. (3.8). For $q \to 0$, most of the vertices in W have degree k. Neglecting the overlapping edges (assuming sparse system), a vertex in the projection therefore will have degree k when the corresponding vertex in Erdős-Rényi has degree $k - w$. Hence:

$$Pr(k) \sim \frac{(np)^{(k-w)} e^{-np}}{(k-w)!}. \tag{3.10}$$

3.3.3 Erdős-Rényi with Barabási-Albert

Barabási-Albert graph $B_{n,m}$ on n vertices is an error-free model to simulate the preferential attachment phenomenon, where m new edges are added at each iteration [10]. Therefore this combination is similar to Barabási-Albert variants with experimental noise [57, 136] in which preferential and random uniform (noise) attachment are combined.

We can apply Fokker-Planck approach to determine the asymptotic behavior. The outline is to begin with an Barabási-Albert graph and iteratively add a uniformly drawn random edge to the graph. The new edges are taken from the non-overlapping edges of Erdős-Rényi. Since there are nm edges in Barabási-Albert, hence there are nmp overlapping edges or $\binom{n}{2}p - nmp$ non-overlapping edges.

Let $u(k,t)$ be the number of vertices of degree k at time step t. At each time step, a new edge will change the degree of 2 vertices. With probability $u(k-1,t)/n$, the number of degree k vertices increases by one if the new edge attaches to a degree $k-1$ vertex. Similarly $u(k,t)$ decreases by one if the new edge attaches to a degree k vertex. Thus

$$u(k,t+1) = u(k,t) + u(k-1,t)/n - u(k,t)/n. \tag{3.11}$$

By replacing t and k by continuous variables, we obtain a partial differential equation which will be a good approximation for large values of k.

$$\frac{\partial u}{\partial t} + \frac{1}{n}\frac{\partial u}{\partial k} = 0. \tag{3.12}$$

The initial condition is the degree distribution of Barabási-Albert, i.e. $u(k,0) \sim 2nm^2k^{-3}$. Hence solve $u(k,t)$ at $t = T = 2 \cdot (\binom{n}{2}p - nmp)$ (twice the number of non-overlapping edges because there are two vertices that change at each time step) to find the degree distribution of the projection:

$$Pr(k) = u(k,T)/n = \frac{2m^2}{(k + 2n(p-p^2))^3}. \tag{3.13}$$

Fig. 3.1 compares the asymptotic expression againsts the analytic expression derived by iterating Eq. (3.11) and the simulations. The continuum approximation of Eq. (3.12)

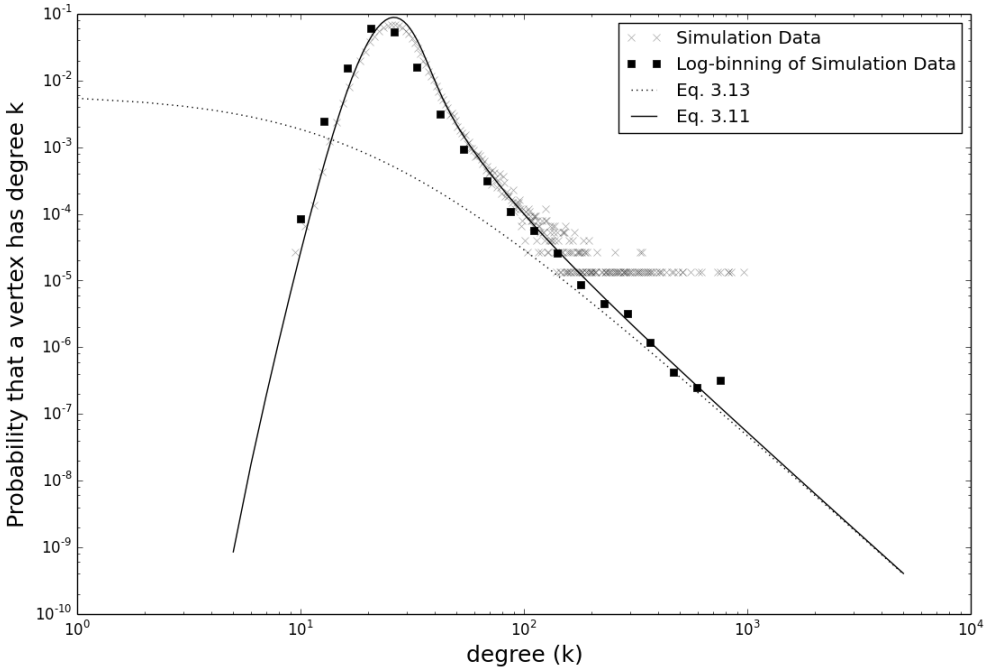

Figure 3.1: The degree distribution of the projection of a Erdős-Rényi graph with a Barabási-Albert graph. The solid line plots the iterative method, Eq. (3.11). The dotted line plots the closed form Eq. (3.13). The crosses represent the distribution obtained from simulations. The degree-distribution log-binning (base 10) of the crosses is plotted with squares.

does not really have a region of validity since the finite number of vertices limits the size of the degree and the asymptotic limit cannot be reach. However, the iterated solution of the Fokker-Planck equation (3.11) matches the simulations.

For the intersection of this combination, if vertex v has degree k in Barabási-Albert graph, then the number of overlapping edges incident to it is $\approx kp$. Hence to derive the degree distribution of this intersection, $Pr^H(deg = x)$, group all the vertices in Barabási-Albert that are most likely to have degree x after the intersection, i.e. $\lfloor kp \rfloor = x$. Thus:

$$Pr^H(deg = \lfloor kp \rfloor) \approx \sum_{i=0}^{\lfloor 1/p \rfloor} Pr^B(deg = \lfloor kp \rfloor + i), \tag{3.14}$$

or

$$Pr^H(deg = x) \approx \sum_{i=0}^{\lfloor 1/p \rfloor} Pr^B(deg = \lceil x/p \rceil + i), \tag{3.15}$$

where $Pr^B(deg = k) \sim k^{-3}$ is the degree distribution of Barabási-Albert graph. Lastly $Pr^B(deg = \lceil x/p \rceil + i) \sim \lceil x/p \rceil^{-3}$ as $x/p \gg i$ for large x. Hence $Pr^H(deg = x) \approx \lfloor 1/p \rfloor \cdot \lceil x/p \rceil^{-3}$, implying that the subgraph is also scale-free (Fig. 3.2).

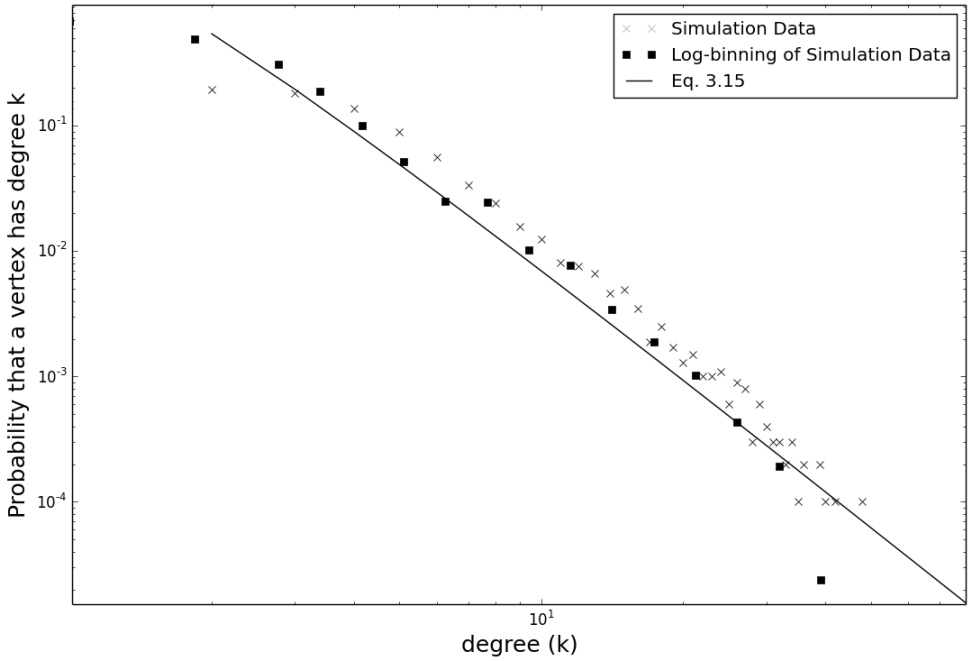

Figure 3.2: The degree distribution of the intersection of Erdős-Rényi and Barabási-Albert.

3.3.4 Watts-Strogatz with Watts-Strogatz

Let $W^1_{n,w,q}$ and $W^2_{n,w',q'}$ be two Watts-Strogatz graphs. If $q \approx q' \to 1$, then the graphs can be expressed as two Erdős-Rényi graphs and their projection was described in section 3.3.1. In contrast the limit $q \to 1$ and $q' \to 0$ is identical to the union of Erdős-Rényi graph and Watts-Strogatz graph in section 3.3.2. Lastly for $q \approx q' \to 0$, both graphs are almost regular hence the degree distribution follows a hypergeometric function $Pr(k) \sim \mathcal{H}(k; n, w, w')$.

3.3.5 Barabási-Albert with Barabási-Albert

If power-law distribution are prevalent in real world systems, then it will be curious and interesting to see the projection of scale-free systems. That is, to consider the projection of the multiplex $\mathcal{M} = \{B^1_{n,m}, B^2_{n,m'}\}$.

The key feature of Barabási-Albert is that the degree distribution follows a power-law function. Hence we want to know if the projection retain this characteristic. Let $Pr^B(k) \sim 2nm^2k^{-3}$ be the probability density function of Barabási-Albert, and to account for the ϵ number of overlapping edges:

1. Probability that a vertex in B^1, v^1 has degree j, i.e. $Pr^B(j)$;
2. Probability that a vertex in B^2, v^2 has degree $k + \epsilon - j$, i.e. $Pr^B(k + \epsilon - j)$;
3. Probability Pr_o that there are ϵ overlapping edges between v^1 and v^2.

Then the combined probability is the given convoluted expression:

$$Pr(k) \sim \sum_{\epsilon}^{k} \sum_{j}^{k} Pr_o(\epsilon | j, k + \epsilon - j) Pr^B(j) Pr^B(k + \epsilon - j) \tag{3.16}$$

where $Pr_o(\epsilon) = \mathcal{H}(\epsilon; n - 1, d_1, d_2)$.

Unfortunately, the asymptotic behavior of Eq. 3.16 is unclear. Simulations (Fig. 3.3) indicate a heavy-tailed distribution, but it is insufficient to suggests that the degree distribution of the projection follows a power-law or log-normal distribution.

3.3.6 Watts-Strogatz with Barabási-Albert

Let $W_{n,w,q}$ and $B_{n,m}$ be Watts-Strogatz and Barabási-Albert graphs respectively, and for the graphs to have equal number of edges, $w \approx 2m$. When $q \to 1$, Watts-Strogatz evolves to an Erdős-Rényi graph and hence the results will be similar to the projection of Erdős-Rényi and Barabási-Albert graphs in section 3.3.3.

For $q \to 0$, since the probability of rewiring is low, most of the lattice edges in Watts-Strogatz remain unchanged. Thus most of the vertices in Watts-Strogatz have degree $\approx w$,

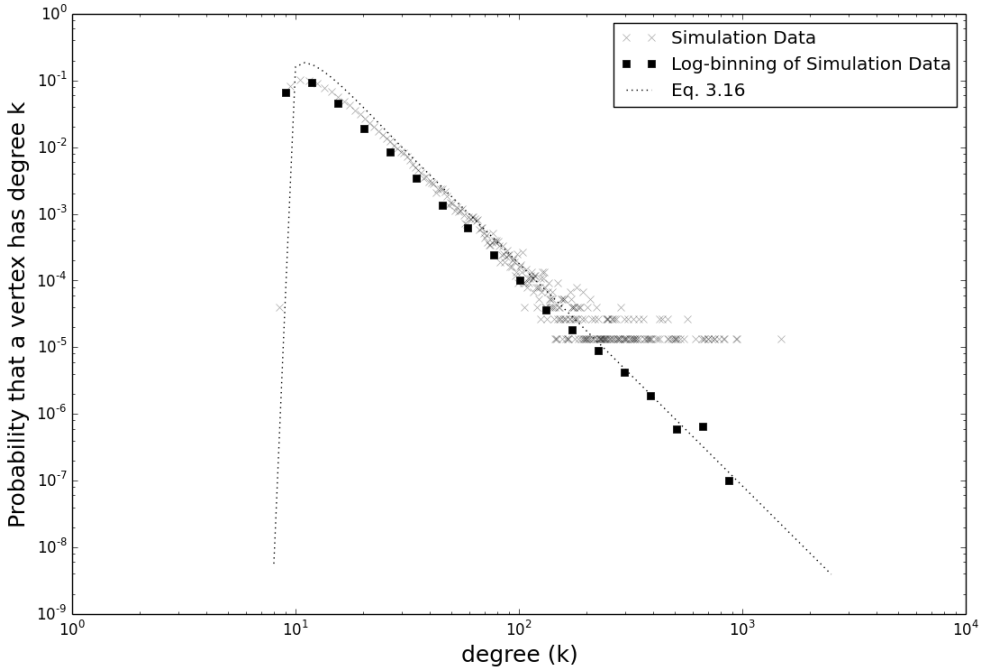

Figure 3.3: The degree distribution of the projection of two Barabási-Albert graphs.

and hence contributes to an increase in the degree of the vertices of Baraási-Albert graph by w. This gives the approximation for the asymptotic behavior (Fig. 3.4):

$$Pr(k) \sim 2m^2(k-w)^{-3}. \tag{3.17}$$

The degree distribution can be refined by considering the overlapping edge. However the expression is in open form and provides little insights, similar to Eq. 3.16. Moreover as k gets significantly greater than w, the overlapping edges at the high degree vertices of Barabási-Albert graphs will be negligible. This implies that the behavior of the projection follows a Barabási-Albert graph for large k, i.e. $Pr(k) \sim 2m^2(k-w)^{-3} \sim 2m^2(k)^{-3}$.

However we are not able to analytically determine the degree distribution of the intersection of Watts-Strogatz and Barabási-Albert graphs, although simulations suggests a power-law-like degree distribution (Fig. 3.4).

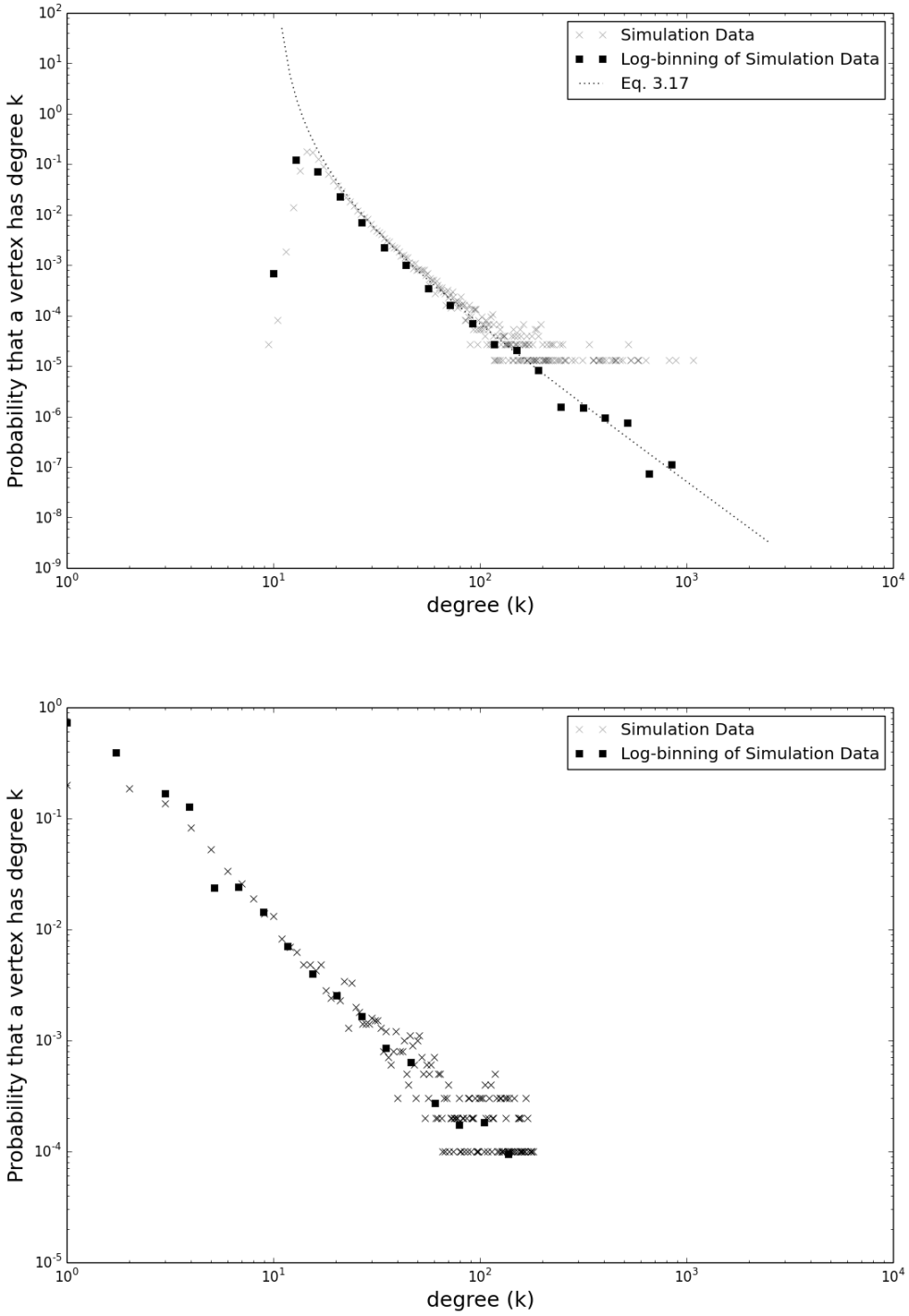

Figure 3.4: Top: The degree distribution of the projection of a Watts-Strogatz and Barabási-Albert graph. Bottom: The degree distribution of the intersection of a Watts-Strogatz and Barabási-Albert graph.

3.4 Clustering Coefficient

Definition 3.4.1 *(**Clustering coefficient** [164]) The local clustering coefficient C_i of vertex v_i is the number of triangles incident to v_i to the number of neighboring pairs of v_i. Let E be the edge set of the graph, and N_i be the set of neighboring vertices of v_i, then:*

$$\textit{clustering coefficient of a graph} = \frac{1}{n}\sum C_i, \tag{3.18}$$

where

$$C_i = \frac{2|\{(v_j, v_k) : v_j, v_k \in N_i, (v_j, v_k) \in E\}|}{deg(v_i)(deg(v_i) - 1)}. \tag{3.19}$$

The clustering coefficient measures how likely the vertices tend to cluster. Similar to the degree distribution of the projection, the distribution of the overlapping triangles affects the accuracy of the clustering coefficient of the projection. However for sparse multiplexes, the number of overlapping triangles can be taken to be negligible (lemma 3.1.6).

3.4.1 Watts-Strogatz with Barabási-Albert

Let $v_{i,P}$ be the i^{th} vertex of the projection graph, and $T(v)$ be the number of triangles at vertex v. The clustering coefficient of the projection is given by:

$$\begin{aligned} \text{Clustering Coefficient of } G^P &= \frac{1}{n}\sum_{i=0}^{n} \text{Clustering Coefficient of vertex } v_{i,P} \\ &= \frac{1}{n}\sum_{i=0}^{n} \frac{T(v_{i,P})}{\binom{deg(v_{i,P})}{2}}. \end{aligned} \tag{3.20}$$

Although there is no approximation to the number of triangles in a Barabási-Albert graph, it is usually very small. Thus $T(v_{i,P}) \approx T(v_{i,ws}) + T(v_{i,ba})$ where $v_{i,ws}$ and $v_{i,ba}$ are the i^{th} vertex of Watts-Strogatz and Barabási-Albert respectively, and hence

$$\begin{aligned} \text{Clustering Coefficient of } G^P &\approx \frac{1}{n}\sum_{i=0}^{n} \frac{T(v_{i,ws}) + T(v_{i,ba})}{\binom{deg(v_{i,c})}{2}} \\ &\geq \frac{1}{n}\sum_{i=0}^{n} \frac{T(v_{i,ws})}{\binom{deg(v_{i,c})}{2}}. \end{aligned} \tag{3.21}$$

Since we are considering $q \to 0$, we can make the following observations: 1) The number of triangles generated by Watts-Strogatz is much more than Barabási-Albert, hence choosing $T(v_{i,ws})$ will get a tighter bound. 2) Most of the vertices in Watts-Strogatz have the same number of triangles attached due to the assumed low rewiring probability. Let τ be the average number of triangles for any vertex in Watts-Strogatz given q. Thus:

$$\begin{aligned}
\text{Clustering Coefficient of } G^P &\geq \frac{1}{n}\sum_{i=0}^{n}\frac{T(v_{i,ws})}{\binom{deg(v_{i,P})}{2}} \\
&\approx \frac{\tau}{n}\sum_{i=0}^{n}\frac{1}{\binom{deg(v_{i,P})}{2}} \\
&= \frac{\tau}{n}\Big(\frac{1}{\binom{2}{2}}+\cdots+\frac{1}{\binom{2}{2}}+\cdots+ \\
&\quad \frac{1}{\binom{k}{2}}+\cdots+\frac{1}{\binom{k}{2}}+\cdots+ \\
&\quad \frac{1}{\binom{n}{2}}+\cdots+\frac{1}{\binom{n}{2}}\Big) \\
&= \tau\Big(\frac{Pr(2)}{\binom{2}{2}}+\cdots+\frac{Pr(k)}{\binom{k}{2}}+\cdots+\frac{Pr(n)}{\binom{n}{2}}\Big) \\
&= \tau\sum_{k=2}^{n}Pr(k)\frac{1}{\binom{k}{2}}, \qquad (3.22)
\end{aligned}$$

where $Pr(k)$ is the degree distribution of the projection graph (section 3.3.6).

From [11], the clustering coefficient of Watts-Strogatz decreases at the rate $(1-q)^3$ and $\tau \approx (w/2)^2$ for $q = 0$. This is also equal to the rate of decrease to the number of triangles since $(1-q)^3$ is the probability that none of the edges of a triangle is rewired. Thus we have the following lower bound on the clustering coefficient of the projection:

$$\text{Clustering Coefficient of } G^P \geq \frac{(1-q)^3 w^2}{4}\sum_{k=0}^{n}P(k)\frac{1}{\binom{k}{2}}. \qquad (3.23)$$

To obtain an upper bound on the clustering coefficient, we can round up the clustering

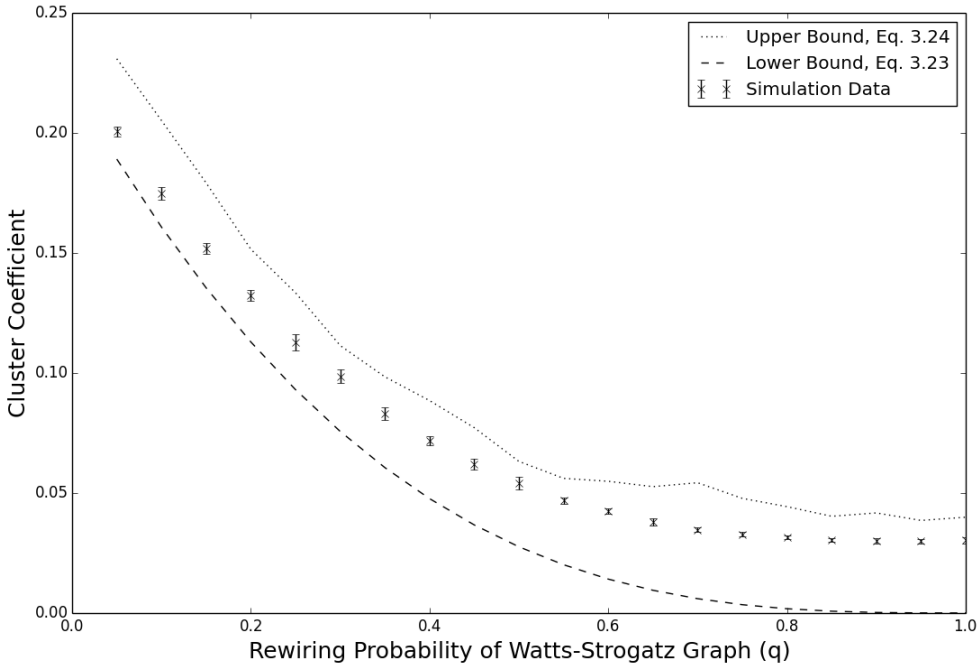

Figure 3.5: The clustering coefficient of the projection of Watts-Strogatz and Barabási-Albert graphs. The standard deviation error bar shows that the exact results are well within the analytical bounds.

coefficient contribution originating from Barabási-Albert:

$$\begin{aligned}
\text{Clustering Coefficient of } G^P &\approx \frac{1}{n}\sum_{i=0}^{n}\frac{T(v_{i,ws})+T(v_{i,ba})}{\binom{deg(v_{i,P})}{2}} \\
&= \frac{1}{n}\sum_{i=0}^{n}\frac{T(v_{i,ws})}{\binom{deg(v_{i,P})}{2}} + \frac{1}{n}\sum_{i=0}^{n}\frac{T(v_{i,ba})}{\binom{deg(v_{i,P})}{2}} \\
&\leq \frac{1}{n}\sum_{i=0}^{n}\frac{T(v_{i,ws})}{\binom{deg(v_{i,P})}{2}} + \frac{1}{n}\sum_{i=0}^{n}\frac{T(v_{i,ba})}{\binom{deg(v_{i,ba})}{2}} \\
&\approx \frac{(1-q)^3w^2}{4}\sum_{k=0}^{n}P^{G^P}(k)\frac{1}{\binom{k}{2}} \\
&\quad + \text{Clustering Coefficient of } B_m. \qquad (3.24)
\end{aligned}$$

3.4.2 Watts-Strogatz with Watts-Strogatz

Similar to the projection of two Barabási-Albert graphs, it is equally interesting to study the projection of two Watt-Strogatz graphs $W^1_{n,w,q}$ and $W^2_{n,w,q'}$. Let $v_{i,1}$ and $v_{i,2}$ be the i^{th} vertices of W^1 and W^2 respectively, which are mapped to same the i^{th} vertex ($v_{i,P}$) in the projection graph.

In the limit $q, q' \to 0$ and $q \approx q'$, almost every vertex of W^1 and W^2 are similar in structure. i.e. $T(v_{i,1}) \approx T(v_{i,2})$ and $deg(v_{i,1}) \approx deg(v_{i,2})$ Hence even with random pairing, every vertex of the projection will be similar. Furthermore for small w, there are few overlapping edges. Hence $T(v_{i,P}) = T(v_{i,1}) + T(v_{i,2}) \approx 2T(v_{i,1})$ and $deg(v_{i,P}) \approx 2deg(v_{i,1})$, then (Fig. 3.6):

$$\begin{aligned} \text{Clustering Coefficient of } G^P &\approx \frac{1}{n}\sum_{i=1}^{n}\left(\frac{2T(v_{i,1})}{\binom{2deg(v_{i,1})}{2}}\right) \\ &\approx \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}\left(\frac{T(v_{i,1})}{\binom{deg(v_{i,1})}{2}}\right) \\ &\approx \frac{1}{2}\text{ Clustering Coefficient of } W^1. \end{aligned} \tag{3.25}$$

3.4.3 Watts-Strogatz with Erdős-Rényi

The clustering coefficient of the projection of Watts-Strogatz with Erdős-Rényi can be estimated in the same way as the previous sections. In the projection the maximum number of triangles at vertex $v_{i,P}$ is $\binom{deg(v_{i,P})}{2}$. Among them, $T(v_{i,ws})$ triangles are from the Watts-Strogatz graph. The rest of the triangles exist if the edges of Erdős-Rényi connect the neighbors of $v_{i,P}$ that are not connected from the edges of Watts-Strogatz.

With probability p from Erdős-Rényi, the number of neighbors pairs at $v_{i,P}$ that are connected by the edges of Erdős-Rényi and not from Watts-Strogatz is:

$$\left(\binom{deg(v_{i,P})}{2} - T(v_{i,ws})\right)p. \tag{3.26}$$

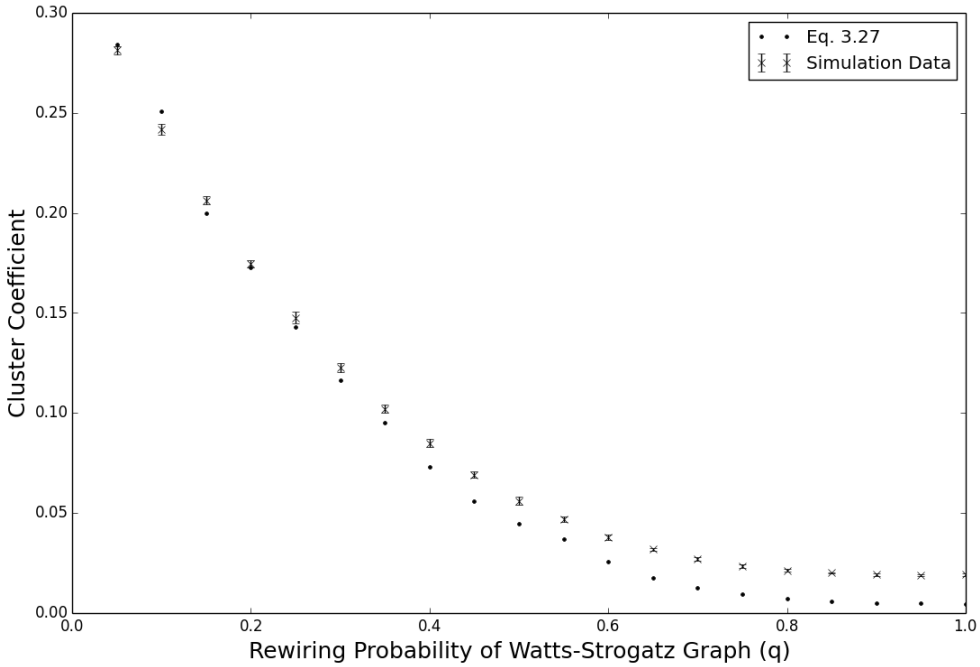

Figure 3.6: The clustering coefficient of the projection of two Watts-Strogatz graphs. Note that as q gets further away from zero, the analytical estimate is many standard deviations away from the empirical results. The reason is that Eq. 3.25 is applicable only for q that is close to zero. For $q \to 1$, the analytical bounds is similar to the projection of two Erdős-Rényi graphs (section 3.3.1).

Hence together with similar approximation from Eq. 3.22 and Eq. 3.23,

$$\begin{aligned}
\text{Clustering Coefficient of } G^P &\approx \frac{1}{n}\sum_{i=1}^{n}\frac{\left(\binom{deg(v_{i,P})}{2} - T(v_{i,ws})\right)p + T(v_{i,ws})}{\binom{deg(v_{i,P})}{2}} \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{\binom{deg(v_{i,P})}{2}p + T(v_{i,ws})(1-p)}{\binom{deg(v_{i,P})}{2}} \\
&= p + \frac{(1-p)}{n}\sum_{i=0}^{n}\frac{T(v_{i,ws})}{\binom{deg(v_{i,P})}{2}} \\
&\approx p + \frac{(1-p)(1-q)^3w^2}{4}\sum_{k=1}^{n}Pr(k)\frac{1}{\binom{k}{2}}, \qquad (3.27)
\end{aligned}$$

where $Pr(k)$ is the degree distribution of the projection graph.

3.4.4 General Observation

There is no analytical result to determine the clustering coefficient of the other combinations of graph ensembles, although it can be numerically observed that their clustering coefficients are generally low (Fig. 3.7). However it is possible to rank the projection of the different combinations according to their clustering coefficients.

Let $^{a}/_{b}$, $^{c}/_{d}$ and $^{e}/_{f}$ be the clustering coefficient of the i^{th} vertex of Erdős-Rényi, Barabási-Albert and Watts-Strogatz respectively, where the numerator is the number of triangles and the denominator is the number of triples. In general the relative ranking of their clustering coefficient is given by:

$$\frac{a}{b} < \frac{c}{d} < \frac{e}{f}. \tag{3.28}$$

If the number of overlapping edges is small, then the clustering coefficient of the i^{th} vertex of the projection is the ratio of the total number of triangles to the total number of triples. E.g. the clustering coefficient of the projection of Erdős-Rényi and Barabási-Albert $\approx (a+c)/(b+d)$. By the mediant inequality,

$$\frac{a}{b} < \frac{a+c}{b+d} < \frac{c}{d} < \frac{c+e}{d+f} < \frac{e}{f}. \tag{3.29}$$

Therefore analytically the projection with the lattice-like Watts-Strogatz (e.g. $(c+e)/(d+f)$) yields higher clustering coefficient than other combinations. In fact from simulations the clustering coefficient of $\{$Watts-Strogatz$\cup$Watts-Strogatz$\}$ > $\{$Watts-Strogatz$\cup$ Barabási-Albert$\}$ > $\{$Watts-Strogatz $\cup$ Erdős-Rényi$\}$ (Fig. 3.7).

Lastly since the clustering coefficient of a subgraph is less than the graph itself, we can deduce that their intersection has low clustering coefficient given that the clustering coefficient of Barabási-Albert network is low.

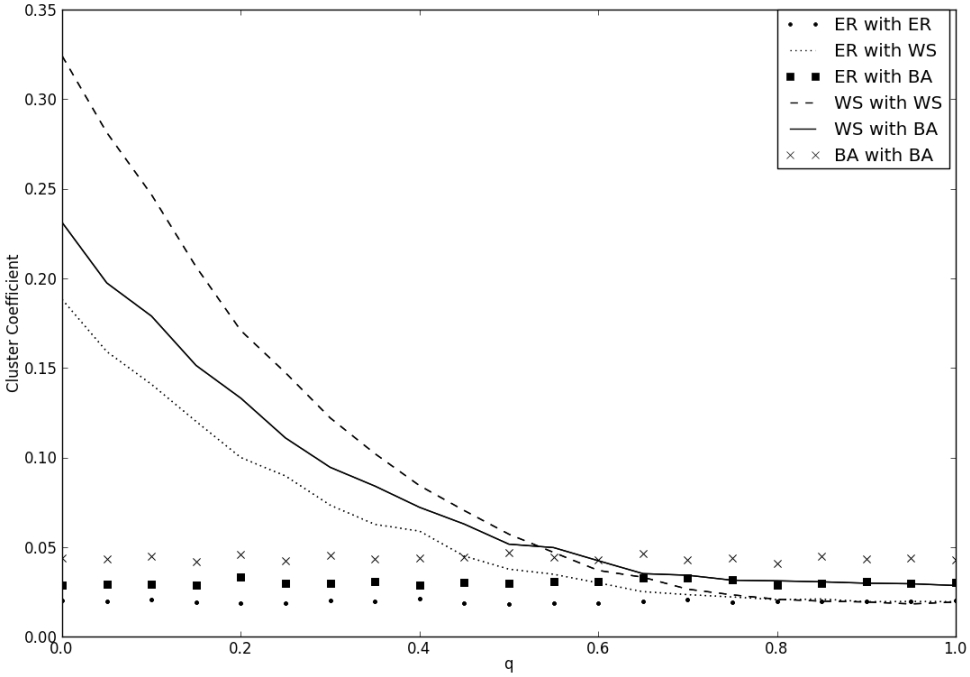

Figure 3.7: Parameters: $n = 1000$, $w = 10$, $m = 5$ and p = $k/(n-1)$. Let ER, WS and BA be the abbreviation of Erdős-Rényi, Watts-Strogatz and Barabási-Albert graph. The x-axis varies the rewiring probability (q) of Watts-Strogatz graph. For small q, combinations with Watts-Strogatz graphs are high. Furthermore, the clustering coefficient of $WS \cup WS$ is greater than $WS \cup BA$, which is greater than $WS \cup ER$

3.5 Centrality

The centrality of a network is the ranking of the vertices according to their relative importance when information flows through the network. A high centrality vertex is known as the hub of the graph, where its absence could severely decrease the efficiency of communication. For example a high centrality vertex could be a major airport, city or a celebrity in a social network. Therefore the interactions between the top centrality vertices of different networks in a multiplex affect the dynamics of the system.

The different centrality metrics like Degree, Betweenness and Eigenvector Centrality are used to reveal the different dynamics of information flow. However they are positively well correlated in general [163], since the mechanics of the other two centrality measures favor vertices with high Degree Centrality. Furthermore Betweenness and Eigenvector Centrality are hard to compute efficiently for large networks, hence Degree Centrality is often used to approximate these centrality rankings instead.

For example Eigenvector Centrality is the stable state of all vertices where a vertex's score is the sum of the centrality scores of its neighbors. Hence vertices with higher degree (equivalently more neighbors) have more components in the sum, and this results in higher Eigenvector Centrality score.

Similarly the Betweenness Centrality of a vertex v is the probability that the shortest path between a randomly chosen vertex pair passes through v. A high degree vertex has many edges leading into it and will therefore be more likely, than a low degree vertex, to connect to a given shortest path between two arbitrarily chosen vertices.

The motivation of this research is that centrality is usually computationally expensive, and it will be helpful to have some theoretical understanding on the stability of centrality ranking. In short to extrapolate some information about the centrality without recomputing the projection graph, specifically the probability that the top centrality vertex v^1_{max} in G^1 remains top in the projection: $Pr(v^1_{max} = v^P_{max})$.

3.5.1 Distribution of Maximum Degree of Erdős-Rényi

Theorem 3.5.1 *Given $G_{n,p}$, where $0 < p < 1$ depends on n. If $\lim_{n\to\infty} np(1-p)/(\ln(n))^3 \to \infty$ and x is a fixed real number, then from [24]:*

$$\lim_{n\to\infty} Pr(\textit{Max Degree} < a + bx) = e^{-e^{-x}},$$

where

$$a = np + \sqrt{2p(1-p)n\ln n}\Big(1 - \frac{\ln\ln n}{4\ln n} - \frac{\ln(2\sqrt{\pi})}{2\ln n}\Big);$$

$$b = \frac{\sqrt{2np(1-p)\ln n}}{2\ln n}.$$

Theorem 3.5.2 *Given $G_{n,p}$, where $0 < p < 1$ depends on n. If $\lim_{n\to\infty} np(1-p)/\ln(n) \to \infty$ and x is a fixed real number, then from [8]:*

$$\lim_{n\to\infty} Pr(\textit{Max Degree} < a + \sqrt{np(1-p)}bx) = e^{-e^{-x}},$$

where a and b defined similarly in Theorem 3.5.1.

These theorems determine the probability that the maximum degree of a random graph is less than bound $\beta_1 = a + bx$ and $\beta_2 = a + \sqrt{np(1-p)}bx$. If we express the expected value of $deg_P(v^1_{max})$ as the same form as those bounds, then we can determine the probability that $deg_P(v^1_{max})$ is the maximum degree of the projection. For clarity Lemma 3.5.3 depends on β_1 from Theorem 3.5.1, and Lemma 3.5.4 depends on β_2 from Theorem 3.5.2.

3.5.2 Stability of The Top Degree Centrality Vertex I

Lemma 3.5.3 *Let $G^1_{n,p}$ and $G^2_{n,q}$ be graphs in a multiplex, where $0 < p < 1$ depends on n. Given v^1_{max} and v^P_{max} be the vertices with the highest Degree Centrality of G^1 and $G^P = G^1 \cup G^2$ respectively. If $deg_1(v^1_{max})$ can be expressed in the same form as β_1 in Theorem 3.5.1, then:*

$$\lim_{n\to\infty} Pr(v^1_{max} = v^P_{max}) \to 0.$$

Proof To simplify the algebra manipulations of $deg_1(v^1_{max}) = a + b\gamma$, we group variable n in each term together, i.e.

$$\mathbb{E}(deg_1(v^1_{max})) = np + \sqrt{p(1-p)}C_1 + \sqrt{p(1-p)}C_2\gamma, \tag{3.30}$$

where

$$C_1 = C_1(n) = \sqrt{2n\ln n}\Big(1 - \frac{\ln\ln n}{4\ln n} - \frac{\ln(2\sqrt{\pi})}{2\ln n}\Big);$$
$$C_2 = C_2(n) = \frac{\sqrt{2n\ln n}}{2\ln n} = \sqrt{\frac{n}{2\ln n}}.$$

The expected number of edges incident at v^1_{max} in G^2 (i.e. $deg_2(v^1_{max})$) is nq, thus the expected number of overlapping edges incident at v^1_{max} is determined by lemma 3.1.4:

$$\begin{aligned} & (np + \sqrt{p(1-p)}C_1 + \sqrt{p(1-p)}C_2\gamma)\cdot(nq)/(n-1) \\ \approx\ & npq + q\sqrt{p(1-p)}C_1 + q\sqrt{p(1-p)}C_2\gamma. \end{aligned} \tag{3.31}$$

Let $p' = 1-(1-p)(1-q)$, the expected degree of v^1_{max} at G^P is Eq. 3.30 + nq - Eq. 3.31:

$$\begin{aligned} \mathbb{E}(deg_c(v^1_{max})) &= (np + \sqrt{p(1-p)}C_1 + \sqrt{p(1-p)}C_2\gamma) + nq \\ &\quad -(npq + q\sqrt{p(1-p)}C_1 + q\sqrt{p(1-p)}C_2\gamma) \\ &= np' + (1-q)\sqrt{p(1-p)}C_1 + \sqrt{p(1-p)}(1-q)C_2\gamma. \end{aligned}$$

We now rearrange the right-hand-side of the expression of $\mathbb{E}(deg_P(v^1_{max}))$ to bring it into the same form as Eq. 3.30, with the probability $p' = 1-(1-p)(1-q)$ instead of p:

$$\begin{aligned} \mathbb{E}(deg_P(v^1_{max})) &= np' + (1-q)\sqrt{p(1-p)}C_1 + (1-q)\sqrt{p(1-p)}C_2\gamma \\ &= np' + \sqrt{p'(1-p')}C_1 + (1-q)\sqrt{p(1-p)}C_2\gamma \\ &\quad -(\sqrt{p'(1-p')} - (1-q)\sqrt{p(1-p)})C_1 \\ &= np' + \sqrt{p'(1-p')}C_1 + \sqrt{p'(1-p')}C_2\Big(\gamma\sqrt{\frac{p(1-q)}{p'}}\Big) \\ &\quad -(\sqrt{p'(1-p')} - (1-q)\sqrt{p(1-p)})C_1 \\ &= np' + \sqrt{p'(1-p')}C_1 + \sqrt{p'(1-p')}C_2x. \end{aligned}$$

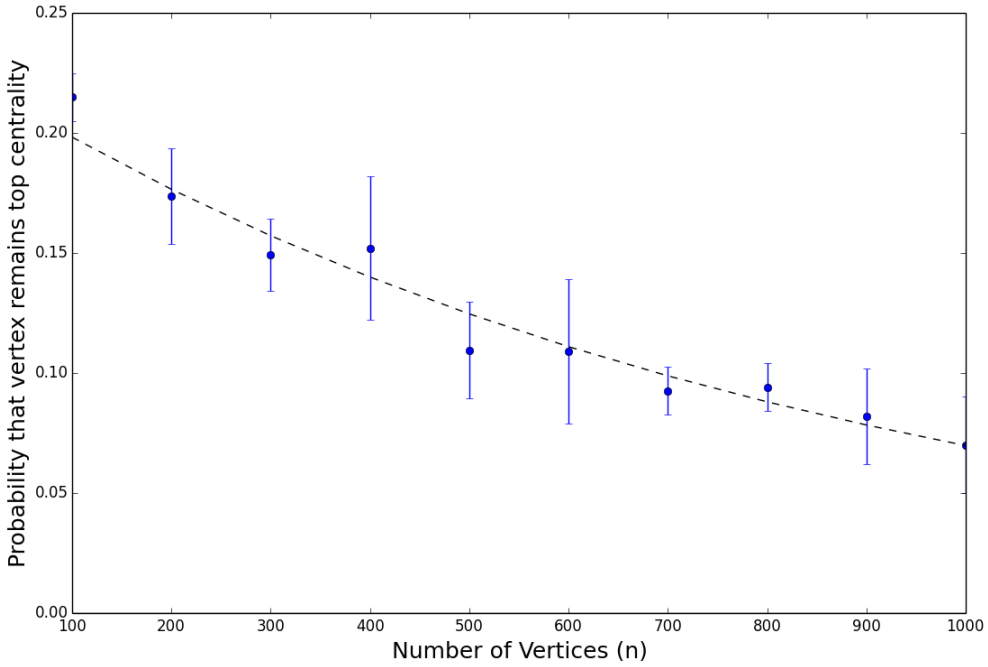

Figure 3.8: The y-axis is the $Pr(v^1_{max} = v^P_{max})$. The best-fit line shows the exponential decay of the empirical simulations of the probability for increasing values of n.

And rewrite the equation such that the rest of the expression is in x:

$$x = \left(\gamma\sqrt{\frac{p(1-q)}{p'}} - \frac{(\sqrt{p'(1-p')} - (1-q)\sqrt{p(1-p)})C_1}{\sqrt{p'(1-p')}C_2}\right) \tag{3.32}$$

$$= \left(\gamma\sqrt{\frac{p(1-q)}{p'}} - \left(1 - \sqrt{\frac{p(1-q)}{p'}}\right)\frac{C_1}{C_2}\right). \tag{3.33}$$

Since $\lim_{n\to\infty} C_1/C_2 = \lim_{n\to\infty} O(\ln n) = \infty$, this implies $x \to -\infty$. The probability that a random graph in the ensemble G^P has maximum degree less than the bound $\mathbb{E}(deg_P(v^1_{max}))$ is given by Theorem 3.5.1, $(exp(-exp(-x)))$ approaches zero.

This result is not particularly surprising since it can be suggested from simulations (Fig. 3.8), where the stability of the top Degree Centrality vertex decreases for increasing values of n. However if we apply the same proof in Lemma 3.5.3 with Theorem 3.5.2 instead of Theorem 3.5.1, then the probability converges to a non-zero limit:

Lemma 3.5.4 *Let $G^1_{n,p}$ and $G^2_{n,q}$ be graphs in a multiplex, where $0 < p < 1$ depends on n. Given v^1_{max} and v^P_{max} be the vertices with the highest Degree Centrality of G^1 and $G^P = G^1 \cup G^2$ respectively. If $deg_1(v^1_{max})$ can be expressed in the same form as β_2 in Theorem 3.5.2, i.e. $a + \sqrt{np(1-p)}b\gamma$ then:*

$$\lim_{n\to\infty} Pr(v^1_{max} = v^P_{max}) \to e^{-e^{-x}},$$

where $x = \frac{\gamma p}{p - pq + q}$.

Proof Similar to the proof of Lemma 3.5.3, we group variable n in each term together:

$$\mathbb{E}(deg_1(v^1_{max})) = np + \sqrt{p(1-p)}C_1(n) + p(1-p)C_2(n)\gamma,$$

where

$$C_1(n) = \sqrt{2n\ln n}\Big(1 - \frac{\ln\ln n}{4\ln n} - \frac{\ln(2\sqrt{\pi})}{2\ln n}\Big);$$

$$C_2(n) = \frac{n\sqrt{2\ln n}}{2\ln n} = \frac{n}{\sqrt{2\ln n}}.$$

With the same chain of algebraic manipulations, we get to the step in Eq. 3.32 where in this case $\lim_{n\to\infty} C_1/C_2 = \lim_{n\to\infty} O(\ln n/\sqrt{n}) = 0$:

$$\begin{aligned} x &= \Big(\frac{\gamma p}{p - pq + q} - \frac{(\sqrt{p'(1-p')} - (1-q)\sqrt{p(1-p)})C_1(n)}{p'(1-p')C_2(n)}\Big) \\ &= \Big(\frac{\gamma p}{p - pq + q}\Big). \end{aligned}$$

Finally since x does not converge to $-\infty$, from Theorem 3.5.2:

$$\lim_{n\to\infty} Pr(\text{Max Degree of } G^P < \mathbb{E}(deg_P(v^1_{max}))) = e^{-e^{-x}} > 0.$$

Lemma 3.5.3 states that if the maximum degree of a G^1 is comparatively "small", then $\lim_{n\to\infty} Pr(v^1_{max} = v^P_{max}) \to 0$. However on the other extreme, if the maximum degree is close to the theoretical maximum i.e. $deg_1(v^1_{max}) \approx n - 1$, then it is likely that it will remain the top degree centrality since no vertices would have degree greater than $n - 1$.

The latter extreme case is uninteresting, but it suggests that there is a phase where the

probability falls between zero and one. Lemma 3.5.4 shows that if an ensemble of graphs whose maximum degree are sufficiently large, they can be expressed in the same form as β_2. Thus the non-zero probability in Lemma 3.5.4 follows.

From experiments the classic Erdős-Rényi model is not apt to demonstrate the non-zero probability via simulations. A random graph picked from a $G_{n,p}$ model tends to fit the conditions of Lemma 3.5.3. If p is small, then the probability that the realization of a graph in $G_{n,p}$ with maximum degree in the same form as β_2 is negligible. However if p is large, the graph will tend to be an almost a complete (and uninteresting) graph.

Therefore it will be more insightful if we parameterize the graphs by their maximum degree and ensure the realization of each graph in the ensemble has equal probability.

3.5.3 Mathematical Properties of $G_{n,\alpha}$

Definition 3.5.5 *Let $G_{n,\alpha}$ be a family of graphs on n vertices. The maximum degree of this ensemble is exactly $\lceil \alpha(n-1) \rceil$, where $0 \leq \alpha \leq 1$.*

A $G_{n,\alpha}$ graph is simply a subgraph of a $\lceil \alpha(n-1) \rceil$-regular graph. Hence the first step to algorithmically generate a $G_{n,\alpha}$ is to pick a random $\lceil \alpha(n-1) \rceil$-regular graph. Next we choose a vertex to be the maximum degree vertex of the ensemble and fixed all the edges incident to it. To uniformly pick all the subgraphs induced by the rest of the edges, remove the edges with probability 0.5.

This algorithm reveals the ensemble's relationship with other graph models, which we can use to deduce further properties of $G_{n,\alpha}$. If a vertex and its incident edges are not fixed, then it is possible for the resultant graph to have maximum degree $< \lceil \alpha(n-1) \rceil$. Such graphs are known as *random f-graphs* [9], where they are in the set of all graphs with a bounded maximum degree. Combine with the fact that $G_{n,\alpha}$ is a subgraph of a regular graph, we have the following relationship:

$$\text{regular graph} \subset G_{n,\alpha} \subset f\text{-graph}.$$

Therefore certain properties like the degree distribution of $G_{n,\alpha}$ can be analyzed by the bounds of f-graphs and regular graphs.

For example Koponen found that almost no vertices in a huge random f-graph have degree less than $d-2 = \lceil \alpha(n-1) \rceil - 2$, i.e. the degree distribution is almost a Dirac Delta Function like a $\lceil \alpha(n-1) \rceil$-regular graph [98]. This implies that the degree distribution of a $G_{n,\alpha}$ follows a Dirac Delta Function too.

Another example is that the exact number of d-regular graphs is hard to count, and in some cases an open problem. Since $G_{n,\alpha}$ is a subgraph of a realization of a regular graph $R_{n,d}$, the number of subgraphs allows us to determine $|G_{n,\alpha}|$. Namely consider the following algorithm: after we picked a vertex from n choices to be the maximum degree and fixed the d edges incident to it, the remaining $d(n/2-1)$ edges induce all the subgraphs. Hence given a $d = \lceil \alpha(n-1) \rceil$-regular graph:

$$|G_{n,\alpha}| = n2^{d(n/2-1)}|R_{n,d}|. \tag{3.34}$$

The fact that it is hard to count regular graphs suggests that it is difficult to prescribe a procedure that algorithmically will uniformly generate all random regular graphs. Consequently the original algorithm to pick a $G_{n,\alpha}$ graph is just as hard. Therefore a heuristic algorithm will be useful for our simulations with large values of n and α.

The number of edges in $R_{n,d}$ is $nd/2$, hence its subgraph $G_{n,\alpha=d/(n-1)}$ has edge set of size between $[d, nd/2]$. Therefore the expected size of the edge set is $\frac{1}{2}(\frac{nd}{2} + d) \approx nd/4 = \alpha n(n-1)/4$. The naive method is repeatedly at random to pick a n-vertices graph with $n(n-1)/4$ edges until one accidentally generates a graph with maximum degree $\alpha(n-1)$. However the probability for this to happen is very small and thus highly inefficient.

The heuristic algorithm is to first find a random $(n-1)$-vertices graph with $(n(n-1)/4 - \alpha(n-1))$ edges. Next we add the last vertex with degree $\lceil \alpha(n-1) \rceil$) and check if the maximum degree of the resultant graph is exactly $\lceil \alpha(n-1) \rceil$. If the condition does not hold, then discard the graph and repeat the algorithm.

A further optimization for our heuristic algorithm is to improve the first phase to find a random $(n-1)$-vertices graph with $(n(n-1)/4 - \alpha(n-1))$ edges. This can be done with the $G_{n-1,p}$ model with mean edge set size $\binom{n}{2}p = \alpha n(n-1)/4 - \alpha(n-1)$, which gives us $p \approx \alpha/2$. In short $G_{n,\alpha}$ is approximated with a random $G_{n-1,p=\alpha/2}$ plus a vertex with degree $\alpha(n-1)$.

3.5.4 Stability of The Top Degree Centrality Vertex II

Consider our main problem with the new ensemble: $G^P = G^1_{n,\alpha} \cup G^2_{n,p}$. In the edge union of two graphs G^1 and G^2, neither graphs' edge set is to dominate the process, i.e. $|E_1| = |E_2|$. Since the expected value of $|E_1| = \lceil \alpha n(n-1) \rceil / 4$, we have $p = \alpha/2$ and the following:

Lemma 3.5.6 *Let v^1_{max} and v^c_{max} be the top degree centrality vertices of $G^1_{n,\alpha}$ and G^P respectively, where $0 < \alpha \leq 1$. Then for $n \to \infty$, $Pr(v^1_{max} = v^c_{max}) \to 1$ (Fig. 3.9).*

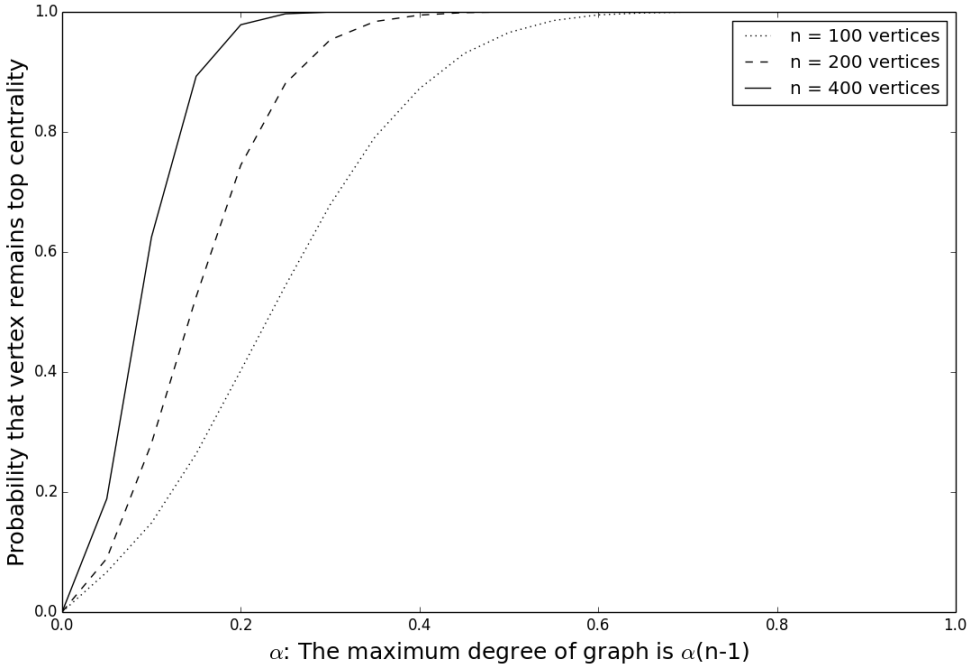

Figure 3.9: Given $0 < \alpha \leq 1$, the y-axis plots the $Pr(v^1_{max} = v^c_{max})$ given the maximum degree of $G^1 = \alpha(n-1)$. For increasing values of n, smaller values of α is required for the probability to be one.

The projection graph can be thought of as being generated from $G^1_{n,\alpha}$, and then follow by adding some random edges in a way given by $G^2_{n,p}$. For G^2 to be significant enough to affect the centrality rankings of G^1, the top degree vertices of G^2 have to be in the same order of magnitude as $\alpha(n-1)$. However these high degree vertices are many standard deviation from the mean (to be shown next), and hence the realization of such G^2 are increasing negligible for $n \to \infty$.

The degree distribution of $G_{n,p}$ is approximately normal with mean and variance $= np$, thus we can estimate the likelihood that a graph in $G_{n,p}$ has a vertex with degree $\alpha(n-1)$ based on the number of standard deviation from the mean. That is if c is the number of standard deviations from the mean, then we express the maximum degree of $G_{n,p}$ as $np + c\sqrt{np}$.

For $G_{n,p}$ to have vertices with degree in the order $\approx \alpha(n-1) = np + c\sqrt{np}$, the number of standard deviations from the mean is $c \approx n(\alpha - p)/\sqrt{np} = \sqrt{\alpha n}$ (since $p = \alpha/2$). Therefore for increasing value of n or α, the realization of such $G_{n,p}$ (and consequently G^2 in Lemma 3.5.6) approaches to zero.

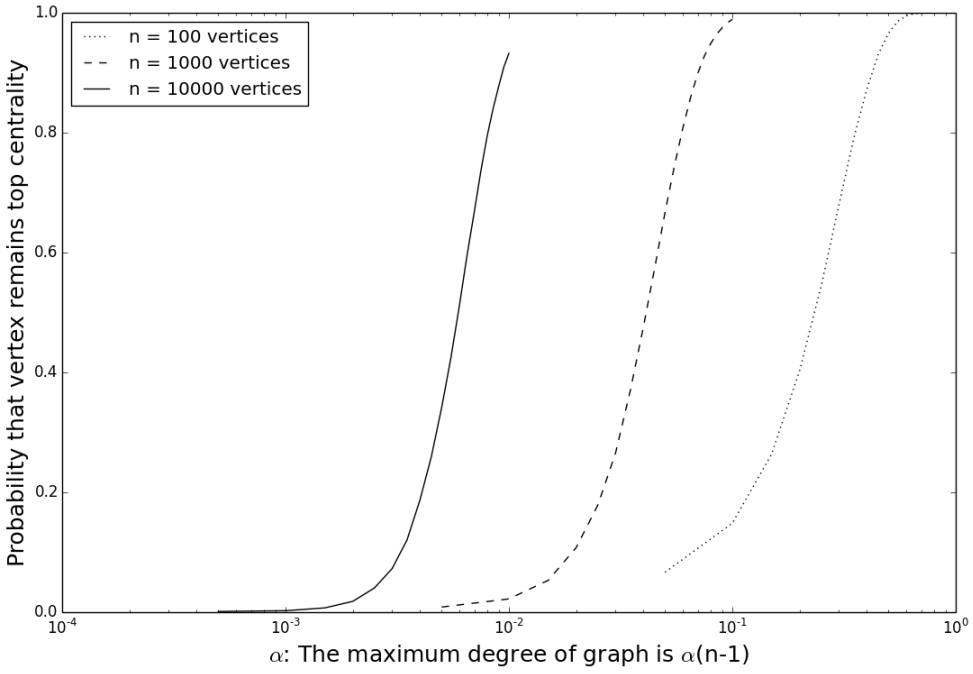

Figure 3.10: Given $0 < \alpha \leq 1$, the y-axis plots the $Pr(v^1_{max} = v^c_{max})$ given the maximum degree of $G^1 = \alpha(n-1)$. The log scale x-axis allows us to see that if α is inversely proportionally to n, then the probability is constant.

In the special case where $\alpha = O(1/n)$, the maximum degree of $G^1_{n,\alpha}$ is a constant standard deviation from the mean of degree of $G^2_{n,p}$. Hence for increasing values of n, there is a constant (and non-negligible) probability for $G^2_{n,p}$ to affect the top centrality ranking of $G^1_{n,\alpha}$ (Fig. 3.10).

Finally to encapsulate the ideas of Lemma 3.5.3 and Lemma 3.5.4, we have to generalize Lemma 3.5.6. Recall that for $v^1_{max} = v^c_{max}$, the $deg_1(v^1_{max})$ at G^1 is many standard deviations c from the mean degree of $G^2_{n,p}$. Since $\lim_{n\to\infty} c \approx n(\alpha - p)/\sqrt{np} \to \infty$, for the probability to be less than one, c has to be a constant given by $(\alpha - p) = O(\sqrt{n})^{-1}$.

If we do not want α or p to depend on n, then only at the threshold $p = \alpha$ where c is a constant. For $p < \alpha$ as shown before, there is always a large enough n such that the $deg_1(v^1_{max})$ at G^1 is many standard deviations c from the mean degree of $G^2_{n,p}$. Conversely for $p > \alpha$, the mean degree of G^2 is $np > n(\alpha)$, which is greater than the maximum degree of G^1. Hence our generalization:

Lemma 3.5.7 *Let v^1_{max} and v^c_{max} are the top degree centrality vertices of $G^1_{n,\alpha}$ and $G^c = G^1_{n,\alpha} \cup G^2_{n,p}$ respectively, where $0 < \alpha, p \leq 1$ are constants. Then for $n \to \infty$,*

$$Pr(v^1_{max} = v^c_{max}) \to \begin{cases} 0 & \text{if } p \geq \alpha, \\ 1 & \text{otherwise.} \end{cases} \tag{3.35}$$

3.6 Connectivity of Evolutionary Interval Graphs

The properties of interval graphs are usually based on the null model where the end points of the intervals were chosen uniformly at random between $[0, 1]$. However this model cannot be parameterized such that the graph is able to evolve from an empty graph to a complete graph. Since connected graphs are usually more interesting, we would like to parameterize the interval graphs such that the hyper-boxes are always connected.

An *evolutionary interval graph* J_r parameterized by variable $r \geq 0$ (known as radius), chooses its intervals' mid-point uniformly at random between $[0, 1]$, with random length between $[0, 2r]$ [150]. It is similar to the Erdős-Rényi graph where increasing r changes the graph from an empty (sparse) graph to a complete (dense) graph. Scheinerman determined the phase transition for the connectivity of evolutionary interval in [150]:

Theorem 3.6.1 *Let J_r be an evolutionary interval graph where the intervals' length are chosen randomly from $[0, 2r]$. If c is a real constant where $r = (\log n + c)/n$, then:*

$$Pr(J_r \text{ is connected}) \to e^{-e^{-c}}. \tag{3.36}$$

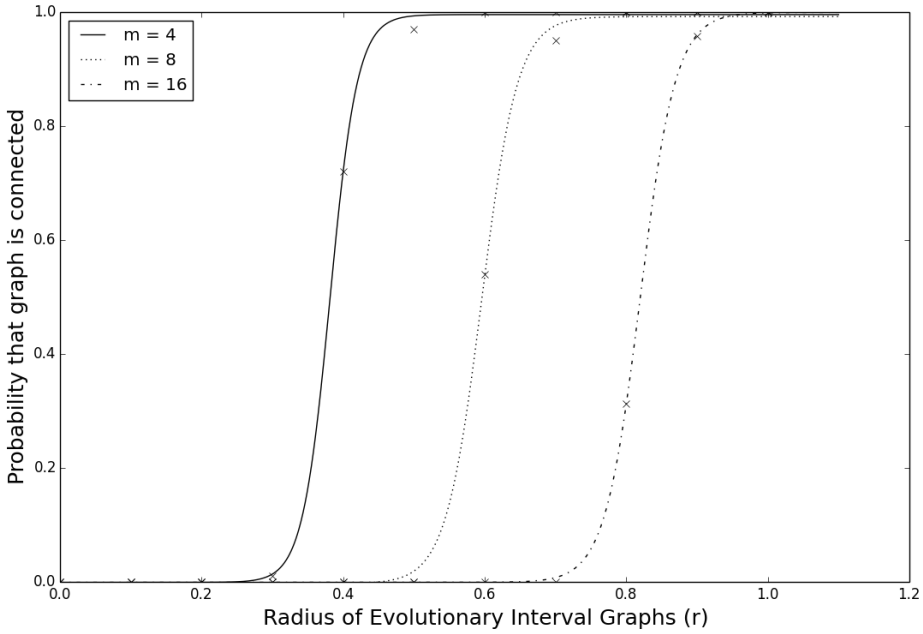

Figure 3.11: The probability that $G = J_r^1 \cap \ldots \cap J_r^m$ is connected for increasing r.

In this thesis we extend Theorem 3.6.1 for the case of the hyper-boxes: Since the edge set of interval graph $E_k \supseteq E$, if G is connected then J_r^k is connected for all k. Thus for G to be connected, it is necessarily (but insufficient) that the set of evolutionary interval graphs are connected (Corollary 3.6.2).

Corollary 3.6.2 *Let graph $G = J_r^1 \cap \ldots \cap J_r^m$, where J_r^k is the k^{th} evolutionary interval graph and its intervals' length are chosen randomly from $[0, 2r]$. If c is a real constant given $r = (\log n + c)/n$, then:*

$$Pr(G \textit{ is connected}) < Pr(J_r^1 \textit{ is connected}) \cdots Pr(J_r^m \textit{ is connected}) \to e^{-me^{-c}}. \quad (3.37)$$

Since $\lim_{m\to\infty} e^{-me^{-c}} \to 0$, it is harder to generate a high dimensional connected graph from a set of evolutionary interval graphs with fixed r. Hence in the experiments we increased r incrementally such that the graph is connected for sufficiently large r (Fig. 3.11).

Chapter 4

Multiplex Communities

Multiplex Communities Detection is a problem to modularize a multiplex into manageable representations. It allows one to either focus on the communities as substructures of a system or the coarse perspective of the system's topology.

Although there are many relevant research on this problem, the foundation is still preliminary. This is due to its disparate nature which causes many researchers to be unaware of the existing research. As a result many of the research are not built upon existing ideas but rather "redesigning" the wheel from scratch.

Therefore the first contribution is the attempt to consolidate the disparate literature on the communities detection for multiplexes. The literature review organized the algorithms and ideas such that there is a system to classify similar concepts.

The second contribution is to compare these algorithms quantitatively with 3 classes of multiplex benchmarks — Random Multiplexes, Structured Multiplexes and Real World Multiplexes. The goal is to illustrate that there are many perspectives of a multiplex community and the wrong choice of algorithm can deviate one from the desired outcome. In fact the Structured Multiplex Benchmark is a proof of example that there can be more than one "ground-truth" solutions in multiplexes.

The research presented in this chapter is published in [111].

4.1 Preliminaries

4.1.1 Introduction to the Communities Detection Problem

The rational to partition a graph or multiplex is to modularize the system into manageable representations. A partition allows one to study the communities as local subsystems and the global interactions between the communities. For instance the user network of a Belgian phone operator can be modularized into 261 communities in which almost everyone in their local community uses a common language. In addition since the communities are weakly tied, the partition also reveals the linguistic split of the Belgian population [22].

Therefore from empirical observations, vertices in the same communities tend to have similar properties [64] and it is usually possible to locally assign meaningful labels to uncharacterized vertices. This has applications to the *link prediction problem* where non-adjacent community members are likely to be connected in the future (homophily) [155].

At the global level, the interactions between the communities is the course-grained perspective of the graph. Thus one should be able to predict the propagation when information flow through the graph, since information tend to trap within the communities [147]. This allows us to apply additional heuristics on NP-hard problems like *Hamiltonian Walk*, that is to find the shortest closed path which visits every vertex at least once [34].

Graph partitioning extends beyond real world applications like VLSI (Very-large-scale integration) circuit designs and task scheduling problem in operation research [4]. It is also used as a strategy in proofs. For example Szemerédi Regularity Lemma [160] is a fundamental tool in extremal graph theory that roughly states that all sufficiently large graphs can be approximated by random-looking graphs. Hence theorems that are easy to prove with random graphs are applicable to sufficiently large graphs [97].

In summary communities detection or graph partitioning is a tool that gives us a different level of abstraction to analyze large graphs/systems. Therefore its applications and research are prevalent in many disciplines including but not limited to computer science, social science, biology and physics. Hence the literature is disparate and there are extensive efforts (including this thesis) to compile the literature on the communities detection problem for graphs (or multiplexes) [64, 137].

4.1.2 Terminologies

As the original communities detection problem is on networks rather than multiplexes, there are many instances in this chapter where the context refers back to the original problem. Hence if there arise ambiguity to the content, we will distinguish the "community" between a multiplex and a monoplex (one of the graphs in the multiplex) as **multiplex-community** and **monoplex-community** respectively. This will avoid confusion when we review the different multiplex communities detection algorithms.

Many of these multiplex-algorithms divide the multiplex problem into independent communities detection problems on the monoplexes. The solutions for these monoplexes are known as **auxiliary-partitions**, and they provide the supplementary information for the multiplex-algorithm to aggregate. The principal solution from the aggregation forms the **multiplex-partition**, which defines the communities in the multiplex.

Unfortunately the term **overlapping** is used in two different contexts in this chapter. The first context refers to the *overlapping edges* as previously defined by Def. 3.1.2. However *overlapping communities* refers to the set of communities where there is at least a pair of distinct communities with common vertices.

4.2 Definitions of a Multiplex-Community

4.2.1 Less Than Ideal Community

In Network Science, a maximal clique is an ideal model of a community as it is not a subset of a larger clique. The community reaches its maximum density of edges and it implies that the community members are tightly connected. However depending on the configuration on the rest of the network, a community might deviate from a clique so as to maximize the modularity of the partition. Even so, the less than ideal (but high quality) community is highly likely to maintain a clique-like structure with high edge density.

Similarly an ideal multiplex-community is a set of vertices where they induce maximal cliques on every graphs in the multiplex. This is also known as *Dense Connected Community Core* [166]. However a less than ideal multiplex-community is harder to conceptualize as it can deviate either in structure or relationship inclusive. For example suppose we

have a multiplex on two relationships, $\mathcal{M} = \{G^1(V, E_1), G^2(V, E_2)\}$ with a multiplex-community on the set of vertices $W \subseteq V$. Let edge sets $E'_1 \subseteq E_1$ and $E'_2 \subseteq E_2$ be the edges among the vertices in W induced by G^1 and G^2 respectively.

A multiplex-community that prioritizes in structural features will regard W as high quality if W is clique-like in G^1 and G^2 simultaneously (section 4.2.3). For instance let E^* be the set of edges such that it forms a clique with W. If E'_1 and E'_2 are random disjoint halves of a clique, i.e. $|E'_1| = |E'_2|$ with $E'_1 \cup E'_2 = E^*$ and $E'_1 \cap E'_2 = \emptyset$, then W is a high quality multiplex-community.

On the other hand, a multiplex-community that prioritizes in relational features will regard W as high quality if there are many overlapping edges (section 4.2.2). In the earlier example, there is no overlapping edges in the multiplex community. Thus a contrasting example is to let the subgraphs induced by W form the same star graphs in G^1 and G^2, such that all the edges overlap i.e. $E'_1 = E'_2$. Although it is clear that star graphs are not clique-like in structure, but some literature define it as a high quality multiplex-community.

In between these two extremes are definitions of multiplex-communities that are in the gray area. The spectrum of definitions on multiplex-communities deviates the quality of a given partition as the multiplex becomes less than ideal (Fig. 4.1). There is a level of qualitative complexity in the decision to define a multiplex-community and hence there is no obvious canonical total ordering to unify these alternative definitions.

4.2.2 Local Definition

From the assumption that a community has weak interactions with the rest of the graph, the evaluation of a community can be isolated or localized. Thus it is possible to establish a community from the perspective of the members in the community.

Consider each graphs in a multiplex as an independent mode of communication between the vertices, e.g. email, telephone, postal, etc. A high quality community should resume high information flow amongst its members when one of the communication modes fails (1 less graph). This is to model communities that demands high reliability like business partners or emergency teams. Hence Berlingerio *et al.* proposed the *redundancy* of the communities as a metric to the quality of a multiplex-community [16]:

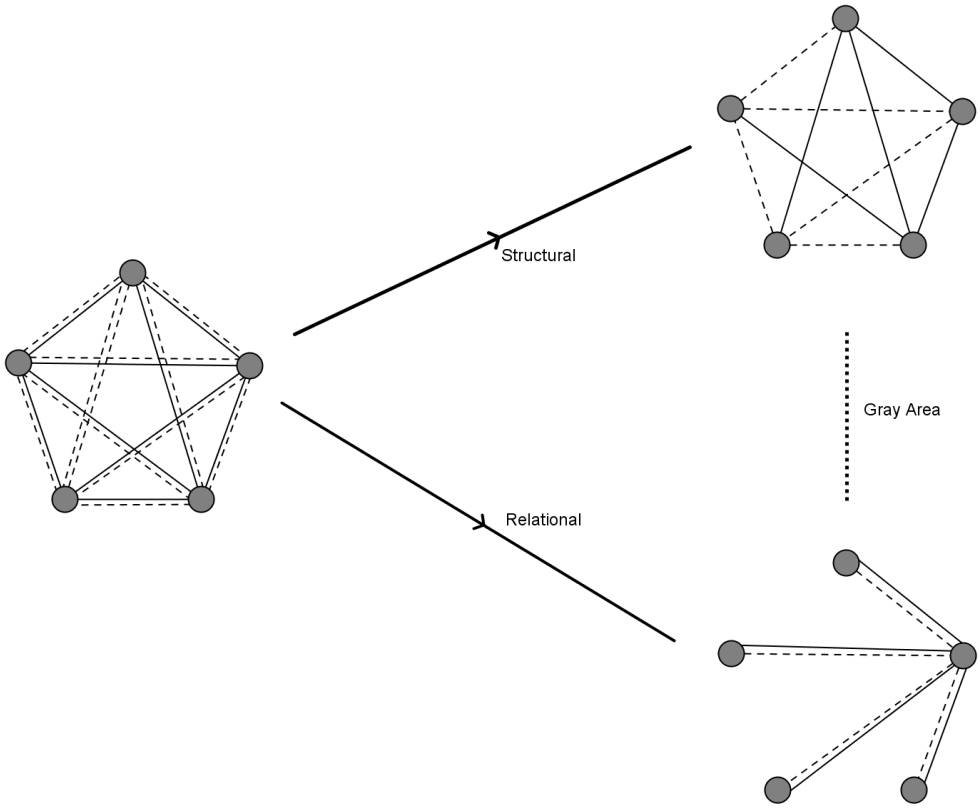

Figure 4.1: Let the multiplex on two relationships be visualized with dotted and solid lines. The figure on the left is an interpretation of an ideal multiplex-community, and the figures on the right are examples of "less-than-ideal" but still high quality multiplex-communities that are based on different definitions in the literature. The top right figure prioritizes on the structural properties and defines a high quality multiplex-community if each graph is clique-like. In contrast figure on the bottom right prioritizes the relational properties of the multiplex and defines a high quality multiplex-community if all the edges overlap. Suppose the problem domain requires the multiplex-community to prioritize on the structural property, but mistakenly uses an algorithm that maximizes the relational properties of a multiplex-community. In that case the results will deviate away from the solution as the multiplex becomes less ideal.

Definition 4.2.1 ***(Redundancy)*** *Given a multiplex $\mathcal{M}$ on vertex set V, a multiplex-community is the subset of vertices $W \subseteq V$ and $R \subseteq W \times W$ be the set of vertex pairs in W that are adjacent in ≥ 1 relationship. The set of redundant vertex pairs are $R' \subseteq R$ where vertex pairs in W are adjacent in ≥ 2 relationships. The redundancy of W is determined by:*

$$\frac{1}{|\mathcal{M}| \times |R|} \sum_{G^i \in \mathcal{M}} \sum_{\{u,v\} \in R'} \delta(u, v, E_i), \tag{4.1}$$

where $\delta(u, v, E_i) = 1$ (zero otherwise) if $\{u, v\} \in E_i$.

Eq. 4.1 counts the number of overlapping edges in the multiplex-community W and the sum is normalized by the maximum possible number of overlapping edges between all adjacent vertex pairs, i.e. $(|\mathcal{M}| \times |R|)$. Roughly speaking the quality of a multiplex-community is a measure of how identical the subgraphs (induced by the vertices of the multiplex-community) are across the graphs in the multiplex.

Thus the number of edges in the multiplex-community is not a necessary condition to its quality. This can lead to an unusual situations where a community is low in density. For instance a cycle of overlapping edges form a community of equal quality (redundancy) as a complete clique of overlapping edges.

4.2.3 Global Definition

The global measure of a partition considers the quality of the communities *and* the interactions among themselves. For instance the *modularity* function (Def. 2.3.2) for monoplex-communities measures how far the communities are from a random graph.

Given a fixed partition on the vertex set, the modularity on each of the m graphs in the multiplex differs. Therefore a good multiplex-community suggests that all the monoplex-communities in the graphs have high modularity. This is analogous to a multi-objective optimization problem where the modularity of each graph is maximized.

To quantify this concept, Tang *et al.* claim that if there exist latent communities in the multiplex, then a subset of the multiplex, $\mathcal{M}' \subset \mathcal{M}$ has sufficient information to determine these communities [161]. For example in a social multiplex, when a group of people is

found to have frequent communications via messaging or email, it is likely that their communication in social media like Facebook is equally strong. If this hypothesis is true, then the communities detected from $\mathcal{M}'$ should reflect high modularity on the rest of the graphs in the multiplex, i.e. $\mathcal{M} \setminus \mathcal{M}'$.

In the language of machine learning, pick a random graph $G \in \mathcal{M}$ as the test data (e.g. Facebook network) and let $\mathcal{M}' = \mathcal{M} \setminus G$ be the training data (e.g. email and messaging network). The multiplex-partition $\mathcal{P}$ yielded from a communities detection algorithm on $\mathcal{M}'$ is evaluated with the modularity function on the test data G. $\mathcal{P}$ is a good multiplex-partition if the modularity of partition $\mathcal{P}$ on the graph G is maximized.

Therefore unlike the local definition, the global perspective of a multiplex-community does not have cases where the community is structurally different from the usual perspective of a monoplex-community. However the global definition have its own set of special cases where it can be contextually unusual for multiplex-communities.

For instance it is possible for a highly modular multiplex-community to have no overlapping edges. This special case is technically still robust from the general idea of "redundancy" in the previous section since the multiplex-community remains modular and tightly connected even after one of the graphs in the multiplex is removed. The issue however is if it is questionable for a multiplex-community to have less than average number of overlapping edges, i.e. pairs in the same community have quantitatively less relationships than with those outside of the community. It concocts a contradicting idea that vertex pairs in the same community have little common relationships.

4.2.4 Vertex Similarity

The concept of homophily is that two vertices belong to the same community if they are similar by some measure. For example the Edge Clustering Coefficient [138] of a vertex pair in a graph measures the (normalized) number of common neighbors between them. A high Edge Clustering Coefficient implies that there are many common neighbors between the vertex pair, thus suggesting that the two vertices should belong in the same community. In the extension for multiplex, Brodka *et. al* introduced Cross-Layer Edge Clustering Coefficient (CLECC) [28].

Definition 4.2.2 ***(Cross-Layer Edge Clustering Coefficient)*** *Given a parameter α, the MIN-Neighbors of vertex v, $N(v, \alpha)$ are the set of vertices that are adjacent to v in at least α graphs. The Cross-Layer Edge Clustering Coefficient of two vertices $u, v \in V$ measures the ratio of their common neighbors to all their neighbors.*

$$CLECC(u, v, \alpha) = \frac{|N(u, \alpha) \cap N(v, \alpha)|}{|N(u, \alpha) \cup N(v, \alpha) \setminus \{u, v\}|}. \tag{4.2}$$

A pair of vertices in a social multiplex with low CLECC suggests that the individuals do not share a common clique of friends through at least α social networks. Therefore it is unlikely that they form a community together. The local and global definition are the extreme ends of the spectrum of the less-than-ideal multiplex-community (Fig. 4.1), and the multiplex-communities defined by vertex similarity tend to fall within the gray areas.

4.3 Theoretical Bounds

The Max-Cut problem finds a partition of a graph such that the number of edges induced across the clusters are maximized. Thus the same partition over the complement graph minimizes the number of edges between the clusters. This is known as the *Balanced-Min-Cut Problem* and it is closely related to the communities detection problem, where the number of edges induced between the communities are minimized.

Therefore to extend our understanding for the communities detection of multiplex, we begin with a known result for the Max-Cut problem. It allows us to prove a corollary for the Balanced-Min-Cut problem on multiplex, which serves as a baseline for the communities detection algorithm.

4.3.1 Maximum Cut Problem on Multiplex

Theorem 4.3.1 *Consider the graphs $G^1, \ldots, G^m$ on the same vertex set V. There exists a k-partition of V into $k \geq 2$ equal-sized communities $C_1, \ldots, C_k$ such that for all $i = 1, \ldots, m$:*

$$\textit{\# edges cut in } G^i \geq \frac{(k-1)|E_i|}{k} - \sqrt{2m|E_i|/k}. \tag{4.3}$$

Proof Sketch [99]: The proof for $k \geq 3$ is similar to $k = 2$, hence we will only prove the latter. WLOG, the vertex set of G^1 is partitioned into 2 equal sets, A and B. Define X_i be an indicator function where $X_i = 1$ if the i^{th} edge is induced from A to B, $X_i = 0$ if otherwise. Since $Pr(X_i) = 1/2$ and with linearity of expectation, we get $\mathbb{E}[X_i] = |E_i|/2$ and $\mathbb{E}[X_i^2] = |E_i||E_i + 1|/4$. By the use of Chebyshev's Inequality, we get the probability that a given partition fail the inequality 4.3 for G^1. Since the probability sum of all graphs to fail is < 1, thus there exists a partition such that all the graphs fulfill inequality 4.3.

Further developments on the bounds were made by imposing additional conditions where the maximum degree is bounded [99] and in cases where $k = m = 2$ [135]. Since the solution for Max-Cut on graphs is NP-complete, the extension to simultaneously Max-Cut all the graphs in a multiplex is naturally NP-complete too. Thus this also implies that balanced minimum bisection is NP-complete too [67].

4.3.2 Balanced Minimum Cut Problem on Multiplex

Corollary 4.3.2 *Consider the graph $G^1, \ldots, G^m$ on the same vertex set V. There exists a k-partition of V into $k \geq 2$ equal-sized communities $C_1, \ldots, C_k$ (i.e. $|C_i| \approx |C_j|$) such that for all $i = 1, \ldots, m$:*

$$\textit{\# edges cut in } G^i \leq \binom{n}{2} - \Big(\frac{(k-1)|\bar{E}|}{k} - \sqrt{2m|\bar{E}|/k}\Big), \tag{4.4}$$

where $|\bar{E}| = \binom{n}{2} - |E_i|$.

Proof Let $\bar{G}^i$ be the complement graph of G^i, and its edge set is denoted by $\bar{E}_i$. Since the maximum number of edges in a graph is $\binom{n}{2}$, hence $|\bar{E}_i| = \binom{n}{2} - |E_i|$. Apply (Max-Cut) Theorem 4.3.1 on the set of complement graphs $\bar{G}_i$ and substitute $|\bar{E}|_i$ into the result, the expression in the corollary follows. The proof in Theorem 4.3.1 ensures that the communities are equal in size.

Using the same proof, we can improve the result for $k = m = 2$ [135] with

Corollary 4.3.3 *Consider the graphs G^1 and G^2 on the same vertex set V. There exists a partition of V into 2 equal-sized communities C_1 and C_2 such that for all $i = 1, 2$:*

$$\textit{\# edges cut in } G^i \leq \binom{n}{2} - \left(\frac{|\bar{E}|}{2} - \sqrt{|\bar{E}|/2}\right), \tag{4.5}$$

where $|\bar{E}| = \binom{n}{2} - |E_i|$.

A partition that fulfills Eq. 4.4 or Eq. 4.5 is not necessarily a good community defined in Section 4.2, vice versa. However the edges induced between partition classes are often perceived as bottlenecks when information flows through the network/multiplex. They are similar to the bridges between cities and communities. Therefore communities detection algorithms tend to minimize the number of edges between different communities.

Unfortunately none of the algorithms in section 4.4 guarantees communities of equal size, thus there is no reasonable way to measures the quality of the algorithms with Eq. 4.3.2. However in the cases where it can be applied, a solution that is greater than the bound implies that the algorithm performs worse than randomization. This is due to the proof in Theorem 4.3.1.

4.4 Communities Detection Algorithms for Multiplex

The general strategy for existing multiplex communities detection algorithm is to extract features from the multiplex and reduce the problem to a more familiar representation. In solving the reduced representation, the multiplex-communities are then deduced from the auxiliary solutions of the reduced problems.

Therefore many multiplex algorithms rely on existing monoplex-communities detection algorithms to derive the auxiliary-partitions for the interim steps. The choice of algorithms is often independent of their extension for multiplex, and hence any communities detection algorithm in theory can be chosen to solve the interim steps. In this thesis our experiments used Louvain Algorithm (section 2.3.2) to generate the auxiliary-partitions as it is the common choice of algorithm in the literature.

4.4.1 Projection

The naive method is to project the multiplex as a weighted graph and then apply monoplex-communities detection algorithm for weighted graphs as the solution. I.e. let A^i be the adjacency matrix of $G^i \in \mathcal{M}$, then the adjacency matrix of the weighted projection of $\mathcal{M}$ is given by $\bar{A} = \frac{1}{m}\sum_{i=1}^{m} A^i$. We will call this the *Projection-Average* of a multiplex.

It has been independently proposed as a baseline for more sophisticated multiplex algorithms as the performance is often "sub-par"* [16, 92, 128, 146]. In our experiments we will also compare with the unweighted variant, that is the *Projection-Binary* of a multiplex, i.e. $G(V, E_1 \cup \ldots \cup E_m)$.

An alternative weight assignment between vertex pair is to consider the connectivity of their neighbors, where a high ratio of common neighbors implies stronger ties [16]. This is based on the idea that members of the same community tend to interact over the same subset of relations, which was also independently proposed by Brodka *et al.* in Def. 4.2.2 [28]. This alternative will be known as *Projection-Neighbors*.

4.4.2 Consensus Clustering

The previous strategy aggregates the graphs first, and then it performs the communities detection algorithm over the resultant graph. It is a poor strategy as it neglects the rich information of the relationships [161]. Therefore the consensus clustering strategy is to first apply the monoplex-communities detection algorithm on the graphs separately as auxiliary partitions, and then the principal partition (the set of multiplex communities) is derived by aggregating the auxiliary partitions in a meaningful manner.

The key concept behind consensus clustering is to measure the frequency with which two vertices are found in the same auxiliary-communities. Vertices that are frequently in the same monoplex-communities are more likely to be in the same multiplex-community. Therefore the communities detection algorithm on the individual graphs in the multiplex determines the structural properties of multiplex-communities, whereas consensus clustering determines the communities' relational properties.

*The projection is merely a different perspective of a community rather than a bad solution.

Frequent Closed Itemsets Mining

Data-mining tends to find a set of items that occurs frequently together in a series of transactions. For example items like milk, cereal and fruits are frequently bought together as a set in supermarkets based on a series of sales transactions. These sets are known as *itemsets*. Berlingerio *et al.* translate the consensus clustering of the auxiliary-partitions as a data-mining problem to discover the multiplex-communities [18].

The vertices in the multiplex define the $|V|$ *transactions* for the data-mining, and the items are tuples (c, d) where the respective vertex belongs in auxiliary-community c in relationship d. For example suppose vertex v_i belongs to auxiliary-communities c_1, c_5 and c_2 in relationships d_1, d_2 and d_3 respectively, then the i^{th} transaction is the set of items $\{(c_1, d_1), (c_5, d_2), (c_2, d_3)\}$. Finally data-mining methods like *Frequent Closed Itemsets Mining* identifies the frequent itemsets as multiplex-communities.

Members in the same multiplex-community (itemsets) are frequently found in the same auxiliary-communities (transaction). E.g. each vertex is a customer's transaction in a supermarket, and a community is a target market (of size e.g. 1000 customers) that the supermarket wants to discover. The Frequent Closed Itemsets Mining takes the customers' transactions as the auxiliary-solutions to extract the multiplex-communities on at least e.g. 1000 vertices, where each customer's transaction is a subset of the target market's itemsets.

Cluster-based Similarity Partitioning Algorithm

Cluster-based Similarity Partitioning Algorithm [161] averages the number of instances vertex pairs are in the same auxiliary-communities. For example in a multiplex with 5 relationships, if there are 3 instances where vertices v_i and v_j are in the same auxiliary-community, then the similarity value of the vertex pair (v_i, v_j) is $3/5$.

Once the similarity is measured for all the vertex pairs, the principal cluster (partition) is determined with k-means clustering — vertices with the closest similarity at each iteration are grouped together. Therefore vertex pairs that are frequently in the same auxiliary-communities will have high similarity value, and hence more likely to be clustered together in the same principal-community. This is similar to the projection-average except that the principal clustering do not necessarily follow a high modularity structure.

Generalized Canonical Correlations

Each of the auxiliary-partitions maps the vertices as points in a l-dimensional[†] Euclidean space. The points are positioned in a way that the shorter the shortest path between two vertices are, the closer they are in the Euclidean space. One of such mapping can be achieved by concatenating the top eigenvectors of the adjacency matrix. Thus given d graphs in a multiplex, there are d structural feature matrices $S^{(i)}$ of size $l \times n$ where the column in each matrix is the position of a vertex in the l-dimensional Euclidean space.

Tang *et al.* want to aggregate the structural feature matrices to a principal structural feature matrix $\bar{S}$ such that the principal partition can be determined from $\bar{S}$ [161]. The "average" $\bar{S} = \frac{1}{d}\sum_{i=1}^{d} S^{(i)}$ however does not result in a sensible principal structural feature matrix since $S^{(i)}$ and $S^{(j)}$ are independent and not in the same Euclidean space.

A solution to fix this problem is to transform the $S^{(i)}$ such that they are in the same space and thus their "average" is sensible. That is the same vertex in the l different Euclidean spaces are aligned to the same point in a common Euclidean space. Specifically we need a set of linear transformations w_i such that they maximize the pairwise correlations of the $S^{(i)}$, and Generalized Canonical Correlations Analysis is one of such standard statistical tools [94]. This allows us to "average" the structures as:

$$\bar{S} = \frac{1}{d}\sum_{i=1}^{d} S^{(i)} w_i, \tag{4.6}$$

and the principal partition is determined via k-mean clustering of $\bar{S}$.

4.4.3 Bridge Detection

A bridge in a graph refers to an edge with high information flow, like the congested roads between two cities, where the absence of these roads separates the cities into isolated communities. One way to do this is to project the multiplex $\mathcal{M}$ to a weighted network and determine the bridges from the projection. Alternatively one can remove them by the definition of a multiplex-bridge to get the desired partitions.

[†]this dimension is different from the dimensions used in multiplex/hyper-box

In social networks, strong ties (edges) are desirable within the communities and weak ties are the bridges between the communities. Hence to identify weak ties between vertex pairs, Brodka *et. al* proposed CLECC (Eq. 4.2) as a metric.

In each iteration, all connected vertex pairs are recomputed and the pair with the lowest CLECC score will be disconnected in all the graphs. The algorithm halts when the desired number of communities is yielded greedily [28]. This is the same strategy presented by Girvan and Newman, where the bridges of a graph are identified by their betweenness centrality score [70]. In the experiments, we let $\alpha = |\mathcal{M}|/2$ for the CLECC computation.

4.4.4 Tensor Decomposition

Algebraic Graph Theory is a branch of Graph Theory where algebraic methods like linear algebra are used to solve the problems on graph. Hence the natural representation for a multiplex is a 3^{rd}-order tensor (as a multidimensional array) instead of a matrix (2^{nd}-order tensor). The set of m graphs in a multiplex is a set of m $n \times n$ adjacency matrices, which can be represented as a $m \times n \times n$ multidimensional array (tensor) [143]. This allows us to leverage on the existing tensor arithmetics like tensor decomposition.

Tensor decompositions are analogues to the Singular Value Decomposition and Lower-Upper Decomposition in matrices, where they express the tensor into simpler components. For example a PARAFAC tensor decomposition [78] is the rank-k approximation of a tensor $\mathcal{T}$ as a sum of rank-one tensors (vectors $\bar{u}^{(i)}$, $\bar{v}^{(i)}$ and $\bar{w}^{(i)}$), i.e.:

$$\mathcal{T} \approx \sum_{i=1}^{k} \bar{u}^{(i)} \circ \bar{v}^{(i)} \circ \bar{w}^{(i)}. \tag{4.7}$$

where $\bar{a} \circ \bar{b}$ denotes the vector outer product. The components in the i^{th} factor, $\bar{u}^{(i)}$, $\bar{v}^{(i)}$ and $\bar{w}^{(i)}$, suggest that there are strong ties (possibly a cluster/community) between the top few elements in $\bar{u}^{(i)}$ and $\bar{v}^{(i)}$ via the relationship of the top component in $\bar{w}^{(i)}$.

For instance suppose the j^{th} element in $\bar{w}^{(i)}$ is the largest element. This suggests that in the i^{th} community, the top 10 (or any predefined threshold) elements in $\bar{u}^{(i)}$ are in the same cluster as the top 10 elements in $\bar{v}^{(i)}$ via the j^{th} dimension/relationship [58, 105, 120, 134].

4.5 Multiplex Benchmarks

An *Erdős-Rényi Graph* is a graph where vertex pairs are connected with a fixed probability. The random nature of this construction usually does not have any meaningful communities structures in them. Hence it is usually not used as a benchmark graph for Communities Detection Algorithms.

A benchmark graph should be similar to the Girvan and Newman Model [70] where some random edges are induced between a set of dense subgraphs (as communities) to form a single connected component. The set of dense subgraphs acts as the "ground-truth" communities of the graph for Communities Detection Algorithms to discover and hence it is apt as a benchmark. This section's goal is to design similar benchmarks for multiplex.

The main challenge is that there is no well accepted definition of a good multiplex community, and hence there is no methodology for us to construct a benchmark such that it does not favor any of the algorithms. Therefore the objective of the following benchmark graphs is **to study the (dis)similarity between the different multiplex-communities detection algorithms**. This allows us to use a collection of highly uncorrelated algorithms to study different perspectives of a multiplex-community.

4.5.1 Unstructured Synthetic Random Multiplex

The simplest construction of a random multiplex is to generate a set of independent graphs on the same vertex set. However many communities detection algorithms are based on some observations of real world multiplexes and these algorithms do not yield interesting results on such random construction. Such random multiplexes is denoted as *Unstructured Synthetic Random Multiplex* (USRM), and they are analogous to Erdős-Rényi Graphs where one should not find any meaningful communities in them.

For simplicity in the numerical experiments there are only two relationships in the multiplexes such that we can easily generate all six combinations of Erdős-Rényi, Watts Strogatz and Barabási-Albert graphs (Table 4.1). See chapter 3 for their statistical properties.

In the experiments there are 128 vertices[‡] and it is important that the number of edges

[‡]chosen to be similar with Girvan and Newman Model in reference [70]

Name	Graph 1	Graph 2
USRM1	Erdős-Rényi	Erdős-Rényi
USRM2	Erdős-Rényi	Watts Strogatz
USRM3	Erdős-Rényi	Barabási-Albert
USRM4	Watts Strogatz	Watts Strogatz
USRM5	Watts Strogatz	Barabási-Albert
USRM6	Barabási-Albert	Barabási-Albert

Table 4.1: Different combinations of multiplexes

in both graphs are equal so that neither graph dominate the interactions in the multiplex. Furthermore to ensure that the Erdős-Rényi graph is connected with high probability, vertex pairs are connected with probability $= \ln 128/128$. Therefore the number of edges in the Erdős-Rényi graph (as well as the other graphs in the multiplex) is $\approx \binom{128}{2} 2 \ln 128/128$.

For higher dimensional USRMs, we will only consider the combinations of Watts Strogatz and Barabási-Albert graphs as their projections exhibit real-world characteristics like high clustering coefficient and power-law like degree distribution [114]. Thus let USRM-Rd_i refers to a multiplex on d relationships with i Watts Strogatz graphs and $(d - i)$ Barabási-Albert graphs.

4.5.2 Structured Synthetic Random Multiplex

This construction is similar to the Girvan and Newman Model where independently generated communities are connected in a way such that the "ground-truth" communities remains. However since there are different perspectives of a multiplex communities (section 4.2), we want to encapsulate all the ideas into a single multiplex benchmark. This is to illustrate the possibility that there could be multiple solutions to the same system.

Structured Synthetic Random Multiplex (SSRM) is a construction where the different definitions of high quality multiplex-communities are found in different multiplex-partitions. Moreover at the same time each of the different multiplex-partition has to be of "poor" quality from the perspective of the other definitions of multiplex-communities.

We begin by generating some subgraphs that are of high quality with respective to the different definitions of multiplex-communities. Next these multiplex-communities are

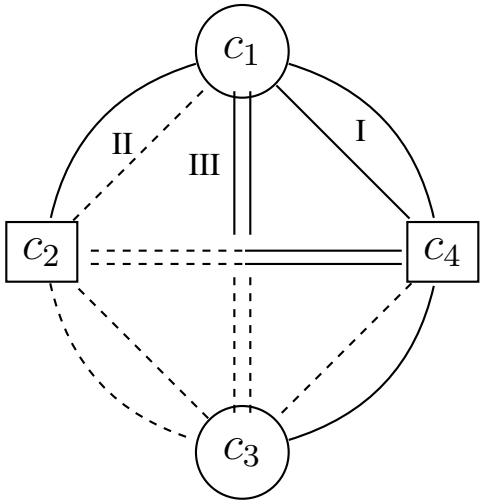

Figure 4.2: To aid visualization, the edges in this SSRM with two relationships are drawn with solid and dashed lines. Let clusters c_1 and c_3 be dense subgraphs where there are more solid edges than dashed edges. Similarly c_2 and c_4 are dense subgraphs with more dashed edges than solid edges. **I:** The solid edges between c_1 & c_4 imply that there are only solid edges between them. This applies the same to the dashed edges between c_2 & c_3. **II:** All the edges between c_1 and c_2 overlaps. **III:** None of the edges between c_1 & c_3 (or c_2 & c_4) overlaps. We denote $\{[c_1, c_2], [c_3, c_4]\}$ as a partition with 2 communities where clusters 1 and 2 form a community and clusters 3 and 4 form the second community. This partition has high redundancy, low modularity and low CLECC multiplex-communities.

modified such that it remains good in one of the three multiplex-communities definitions **and** poor in quality by the other definitions. The final step is to combine these multiplex-communities into a single multiplex as our SSRM benchmark.

Fig. 4.2 shows $\{[c_1, c_2], [c_3, c_4]\}$ as a partition with 2 multiplex-communities where clusters 1 and 2 form the first multiplex-community and clusters 3 and 4 form the second multiplex-community. This partition has high redundancy, but low modularity and low CLECC multiplex-communities. The second partition $\{[c_1, c_3], [c_2, c_4]\}$ has the communities of high modularity, but poor by the rest of the metrics. Lastly $\{[c_2, c_3], [c_1, c_4]\}$ is the final partition where the multiplex-communities have high CLECC vertex pairs. This synthetic multiplex expresses the multi-perspective nature of a multiplex communities.

High Modularity, Low Redundancy & Low CLECC Multiplex-Communities

To create a high modularity multiplex-community, we begin with clusters 1 and cluster 3 as cliques in both relationships (The same construction applies to clusters 2 and 4 respectively). For these clusters to be low in redundancy, remove some edges in both clusters such that there are very few overlapping edges in the clusters while maintaining high modularity. This can be done by making the first graph in both clusters to be the complement of the graph in the second relationship.

The next step is to add random edges between clusters 1 and 3 such that the resultant cluster is a single connected component. We denote $[c_1, c_3]$ as the component connected of cluster 1 and cluster 3. To maintain a low redundancy, the new edges cannot overlap.

Finally tweak the clusters such that the CLECC score is low between many vertex pairs in the combined component of cluster 1 and 3. Specifically the vertex pairs connected by the new edges in the previous step to have low CLECC scores. This is possible if cluster 1 has more edges in the first relationship whereas cluster 3 has more edges in the second relationship. In doing so the neighbors of the vertex in cluster 1 will be significantly different from the neighbors of vertex in cluster 3, thus a low CLECC score.

High CLECC, Low Modularity & Low Redundancy Multiplex-Communities

Since all 4 clusters do not have overlapping edges, the redundancy remains low for any partition on the multiplex. Therefore this section focuses on increasing the CLECC scores. The first step is to make cluster 1 and 4 to be high in CLECC score.

Currently clusters 1 and 4 are similar and hence the neighbors of any given vertex in each cluster is similar too. Therefore by adding new edges between cluster 1 and 4 will not affect the CLECC score. However these new edges should *only* be drawn in the first relationship, since it is the dominant relationship in $[c_1, c_4]$. This simultaneously reduces the modularity of $[c_1, c_4]$ in the second relationship, since the clusters are not connected and the graph in the second relationship is sparse. This gives multiplex-communities $[c_1, c_4]$ a low modularity while maintaining the high CLECC score.

The construction is similar for clusters 2 and 3, with the exception that only edges in the second relationship connects the clusters together.

High Redundancy, Low Modularity & Low CLECC Multiplex-Communities

Given $[c_1, c_3]$ have low CLECC score and that clusters 2 and 3 are similar, the CLECC score for $[c_1, c_2]$ should be similar to $[c_1, c_3]$. Thus this section is to find a way to connect clusters 1 and 2 such that the redundancy of $[c_1, c_2]$ is high. Redundancy is measured by Eq. 4.1, where it a function to the number of overlapping edges. Since there is no overlapping edges at this point of the construction, the redundancy increases when new overlapping edges between clusters to form the multiplex-communities $[c_1, c_2]$ and $[c_3, c_4]$.

Although $[c_1, c_2]$ and $[c_3, c_4]$ have relatively high redundancy as compared to other partition in the multiplex, it is plausible that they have lower redundancy than a random community in USRM1. To nudge the redundancy higher, it is necessary to add new edges such that there are overlaps in the four clusters. However this might increase the modularity of $[c_1, c_2]$ which we want to avoid. Therefore this final step has to be done incrementally.

Evaluation of the different ground truth partitions

$\{[c_1, c_2], [c_3, c_4]\}$, $\{[c_2, c_3], [c_1, c_4]\}$ and $\{[c_1, c_3], [c_2, c_4]\}$ are the different "ground-truth" multiplex-partitions, where each of them represents a different "ideal-partition". However simultaneously they are poor from the perspective of the other metrics.

	Redundancy	*CLECC	**Modularity
$[c_1, c_2]$	**0.0492**	0.1142	-0.0287
$[c_3, c_4]$	**0.0537**	0.1087	-0.0332
$[c_2, c_3]$	0	**0.1541**	0.007
$[c_1, c_4]$	0	**0.1642**	0.012
$[c_1, c_3]$	0	0.1113	**0.0317**
$[c_2, c_4]$	0	0.1083	**0.0245**
Random	0.0217	0.1056	0.0033

Table 4.2: The rows (except the last row) are paired up such that they implies a common partition. E.g. the first two rows are communities of the partition $\{[c_1, c_2], [c_3, c_4]\}$. The redundancy and CLECC score of community $[c_1, c_2]$ are 0.0492 and 0.1142 respectively. ***CLECC:** Average CLECC score between all vertex pairs in the community. ****Modularity:** The two values in the partition refers to the modularity of the two graphs in the multiplex. E.g. partition $\{[c_1, c_2], [c_3, c_4]\}$ has modularity -0.0287 and -0.0332 for the first and second relationships respectively. A partition is "poor" if its measurement is closer to a random partition (last row) than the maximum (values in bold).

For example Table 4.2 shows that the partition $\{[c_1, c_2], [c_3, c_4]\}$ has communities with the maximum redundancy. However it has multi-modularity and CLECC scores similar to a random partition, suggesting that it is poor relative to other metrics. This shows that these metrics are conceptually different.

4.5.3 Real World Multiplex

The issue with synthetic multiplex is that it does not truly reflects the real-world systems. The relationships between vertices are artificially imposed such that we can distinguish the different communities detection algorithms. However real-world communities do not behave in such a systematic manner, therefore we have to compare the multiplex-communities detection algorithms with real-world multiplexes from open dataset.

Youtube Social Network

Youtube is a video sharing website that allows interactions between the video creators and their viewers. Tang *et al.* collected 15,088 active users to form a multiplex with 5 relationships where two users are connected if: 1) they are in the contact list of each other; 2) their contact list overlaps; 3) they subscribe to the same user's channel; 4) they have subscription from a common user; 5) they share common favorite videos [161].

Transportation Network

A multirelational transportation network is known as a multimodal network, where bus stops, train stations and terminals are indistinguishable locations (vertices) to transit to a different mode of transportation. Cardillo *et al.* constructed an air traffic multiplex from the data of European Air Transportation (EAT) Network with 450 airports as vertices [37]. An edge is drawn between two vertices if there is a direct flight between them. Each of the 37 distinct airlines in the EAT Network forms a relationship between the airports.

4.6 Comparing Partitions

Normalized Mutual Information (NMI) [118] is a popular similarity metric for the network partition, with a real-value score between $[0, 1]$ (1 implies identical). However NMI does not measure overlapping communities that are yielded by Communities Detection Algorithms like *Frequent Closed Itemsets Mining* and *Tensor Decomposition*.

Furthermore these algorithms also do not ensure that all vertices belong in at least one community, i.e. "isolated" vertices with no community membership. Therefore even variants of NMI [101] for overlapping communities are not applicable. Hence in such cases we will use the Omega Index [48], a generalized variant of the Adjusted Rand Index [84].

4.6.1 Normalized Mutual Information

Mutual Information $I(\mathcal{A};\mathcal{B})$ measures the information of the communities-membership of all vertex-pairs in A given the communities-membership in B, vice versa. Roughly speaking, given the information on $\mathcal{B}$ how well can we guess that a vertex pair is in the same community in $\mathcal{A}$? Formally this is defined as $I(\mathcal{A};\mathcal{B}) = H(\mathcal{A}) - H(\mathcal{A}|\mathcal{B})$ where $H(\mathcal{A})$ is the Shannon entropy:

$$H(\mathcal{A}) = -\sum_{k} P(a_k) \log P(a_k), \tag{4.8}$$

where $P(a_k)$ is the probability of a random vertex is in the k^{th} community of partition $\mathcal{A}$, i.e. the ratio of the size of the k^{th} community to the total number of vertices. Similarly $P(b_k)$ denotes the case for partition $\mathcal{B}$. Thus the Mutual Information is expressed as:

$$I(\mathcal{A};\mathcal{B}) = \sum_{j}\sum_{k} P(a_k \cap b_j) \log \frac{P(a_k \cap b_j)}{P(a_k)P(b_j)}, \tag{4.9}$$

where $P(a_k \cap b_j)$ is the probability that a random vertex is both in k^{th} and j^{th} communities. Basically the larger the intersection of $\mathcal{A}$ and $\mathcal{B}$ is, the greater the Mutual Information. Finally to normalize the score:

$$\text{NMI}(\mathcal{A},\mathcal{B}) = \frac{I(\mathcal{A};\mathcal{B})}{[H(\mathcal{A}) + H(\mathcal{B})]/2}. \tag{4.10}$$

4.6.2 Omega Index

The unadjusted Omega Index averages the number of vertex pairs that are in the *same number* of communities. Such vertex pairs are known to be in *agreement*. Consider the case with two partitions $\mathcal{A}$ and $\mathcal{B}$, and the number of communities in them are $|\mathcal{A}|$ and $|\mathcal{B}|$ respectively. The function $t_j(\mathcal{A})$ returns the set of vertex pairs that appears exactly in $j \geq 0$ overlapping communities in $\mathcal{A}$. Thus the unadjusted Omega Index:

$$\omega_u(\mathcal{A};\mathcal{B}) = \frac{1}{\binom{n}{2}} \sum_{j=0}^{max(|\mathcal{A}|,|\mathcal{B}|)} |t_j(\mathcal{A}) \cap t_j(\mathcal{B})|. \tag{4.11}$$

To account for vertex pairs that are allocated into the same communities by chance, we have to subtract it from expected omega index of a null model:

$$\omega_e(\mathcal{A};\mathcal{B}) = \frac{1}{\binom{n}{2}^2} \sum_{j=0}^{max(|\mathcal{A}|,|\mathcal{B}|)} |t_j(\mathcal{A})| \cdot |t_j(\mathcal{B})|. \tag{4.12}$$

Finally the Omega Index is given by:

$$\omega(\mathcal{A};\mathcal{B}) = \frac{\omega_u(\mathcal{A};\mathcal{B}) - \omega_e(\mathcal{A};\mathcal{B})}{1 - \omega_e(\mathcal{A};\mathcal{B})}. \tag{4.13}$$

It is possible for the Omega Index to be negative, where there are less agreement than pure stochastic coincidence would expect. It is regarded as uninteresting as it does not suggests anything more than the fact that the partitions are not similar. Identical partitions have Omega Index of 1.

4.6.3 Notations For Empirical Results

To simplify the results and figures from our experiments, we will use some shorthands to denote the algorithms and partitions. For communities detection algorithms, we use $\mathcal{A}$ and $\mathcal{P}$ to denote "Algorithm" and "SSRM Partition" respectively (Table 4.3).

Notation	Description
$\mathcal{A}_1$	Projection-Binary
$\mathcal{A}_2$	Projection-Average
$\mathcal{A}_3$	Projection-Neighbors
$\mathcal{A}_4$	Cluster-based Similarity Partition Algorithm
$\mathcal{A}_5$	Generalized Canonical Correlations
$\mathcal{A}_6$	CLECC Bridge Detection
$\mathcal{A}_7$	Frequent Closed Itemsets Mining
$\mathcal{A}_8$	PARAFAC, Tensor Decomposition
$\mathcal{P}_1$	$\{[c_1, c_2], [c_3, c_4]\}$ of SSRM
$\mathcal{P}_2$	$\{[c_2, c_3], [c_1, c_4]\}$ of SSRM
$\mathcal{P}_3$	$\{[c_1, c_3], [c_2, c_4]\}$ of SSRM

Table 4.3: Shorthands for the different algorithms and "ground-truth" communities.

4.7 Empirical Comparison of the Algorithms

4.7.1 Algorithm Parameters

Some algorithms require additional parameters which are independent to the multiplex, e.g. Frequent Closed Itemsets Mining need to define the minimum community size. In the experiments, the value is chosen such that there are ≥ 2 communities in the solution.

The main parameter for CLECC Bridge Detection is the α in Def. 4.2.2, which the neighbors of a vertex have to be adjacent in at least α graphs. We let $\alpha = \lceil |\mathcal{M}|/2 \rceil$ as there was no consistent value such that it will always yield meaningful communities for all the benchmarks.

PARAFAC is parameterized by the rank of the approximation and the threshold to define the top x components in the rank-one tensors. This is usually done by manually fine-tuning [58, 105, 120, 134], however it is infeasible for our numerous random trials. Thus x is chosen such that the difference between the x^{th} and $x + 1^{th}$ element is greater than the average difference among the elements in the rank-one tensor.

Lastly the final step for Cluster-based Similarity Partition and Generalized Canonical Correlations is k-mean clustering. Since these algorithms maximize the modularity of every graph in the multiplex, the best values of k are chosen such that it maximizes the multi-modularity (section 4.2.3).

4.7.2 Unstructured Synthetic Random Multiplex

Figure 4.3 shows the Omega Index of all the pairwise multiplex-communities detection algorithms for the USRM benchmarks. The last 13 boxplots on the right are the pairwise comparisons with the overlapping-communities algorithms, i.e. $\mathcal{A}_7$ and $\mathcal{A}_8$.

The first observation is that $\mathcal{A}_8$ (PARAFAC) is not similar to any of the algorithms. One of the reasons is that it is hard to choose the right parameters for PARAFAC, i.e the k approximation for Eq. 4.7 and the predefined threshold for the top few elements of the rank-one tensors. There is no systematic method to choose the rank besides manual observation for every given multiplex [58, 105, 120, 134].

However overlapping-communities algorithm $\mathcal{A}_7$ (Frequent Closed Itemsets Mining) is similar to a class of non-overlapping algorithms, i.e. $\mathcal{A}_1$ to $\mathcal{A}_4$. Specifically $\mathcal{A}_i$ is similar to the class of Projection algorithms $\mathcal{A}_1$ to $\mathcal{A}_3$. In fact the Omega Index and NMI scores (Fig. 4.4) for all pairwise comparisons of the algorithm family $\mathcal{F} = \{\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3, \mathcal{A}_4, \mathcal{A}_7\}$ is generally greater than the other algorithm pairs.

The family of algorithms $\mathcal{F}$ has low pairwise Omega Index and NMI scores for USRM 1, 3 and 6. There is no fundamental reason to why the class of projection algorithms ($\mathcal{A}_1$ to $\mathcal{A}_3$) should produce non similar communities, therefore we deduce that USRM 1, 3 and 6 are not good multiplexes for benchmarking.

Since USRM1 is the combination of two Erdős-Rényi graphs, thus naturally there is no community structure. Whereas USRM 3 and 6 are the combinations of Barabási-Albert graph with Erdős-Rényi and Barabási-Albert respectively. Although Barabási-Albert has structural properties, it has low Clustering Coefficient (similar to Erdős-Rényi), which means that the vertices do not have the tendency to cluster together. Therefore the vertices in USRM 3 and 6 do not form communities.

Lastly in higher dimension, the observations are different where Fig. 4.5 shows the Omega Index of various parameters of USRM-Rd_i. Firstly the family of algorithms $\mathcal{F} = \{\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3, \mathcal{A}_4, \mathcal{A}_7\}$ is no longer pairwise similar. Although Projection-Binary ($\mathcal{A}_1$) is somewhat similar to Projection-Average, Projection-Neighbors and Cluster-based Similarity Partition ($\mathcal{A}_2$,$\mathcal{A}_3$,$\mathcal{A}_4$), the rest of the algorithms are not pairwise similar.

The assignment of the weight on the edges for Projection-Average and Projection-

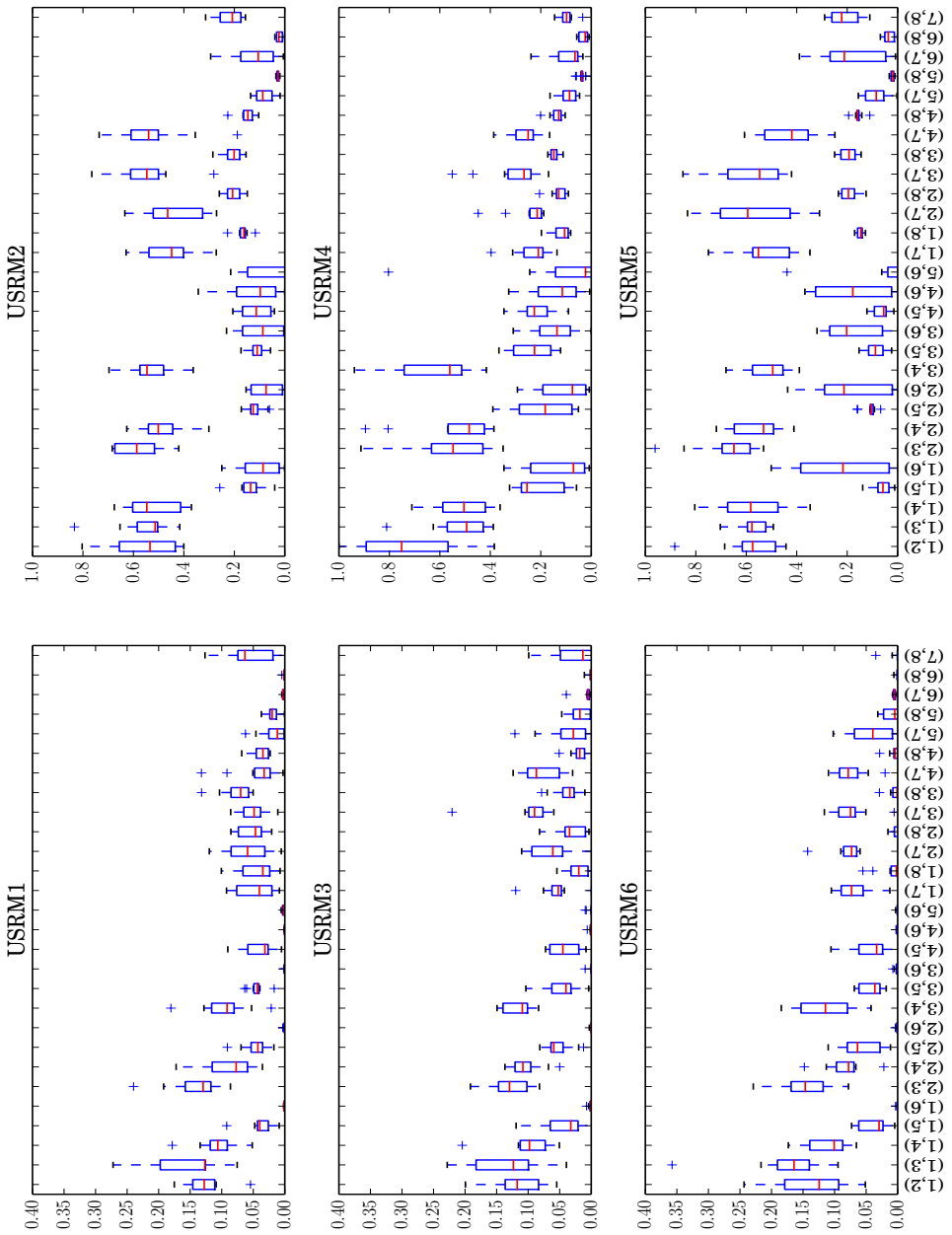

Figure 4.3: The Omega Index of all pairwise multiplex-communities detection algorithms for the different benchmark USRM. The tuple (i, j) on the x-axis refers to the pairwise comparison of $\mathcal{A}_i$ and $\mathcal{A}_j$. The tuples are arranged such that the comparisons with overlapping-communities algorithms (13 tuples) are placed on the right. Note that the scale of the figures on the left is different from the scale on the right.

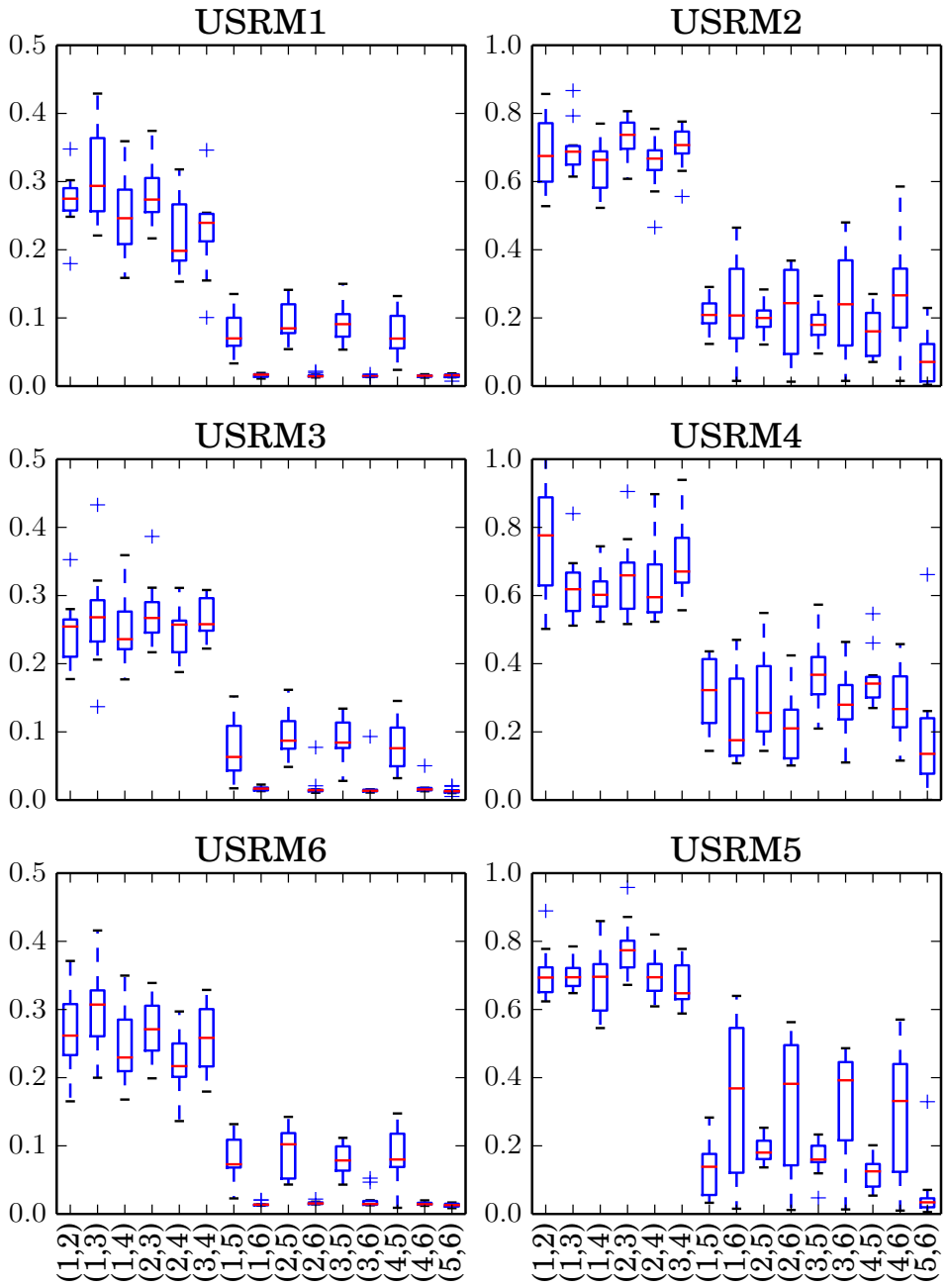

Figure 4.4: The NMI scores of all pairwise non-overlapping multiplex-communities detection algorithms for USRM benchmarks. The tuples are arranged such that pairwise comparisons of $\{\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3, \mathcal{A}_4\}$ are grouped to the left of the boxplots. The "interesting" figures USRM 2, 4 and 5 are placed to the right for comparison.

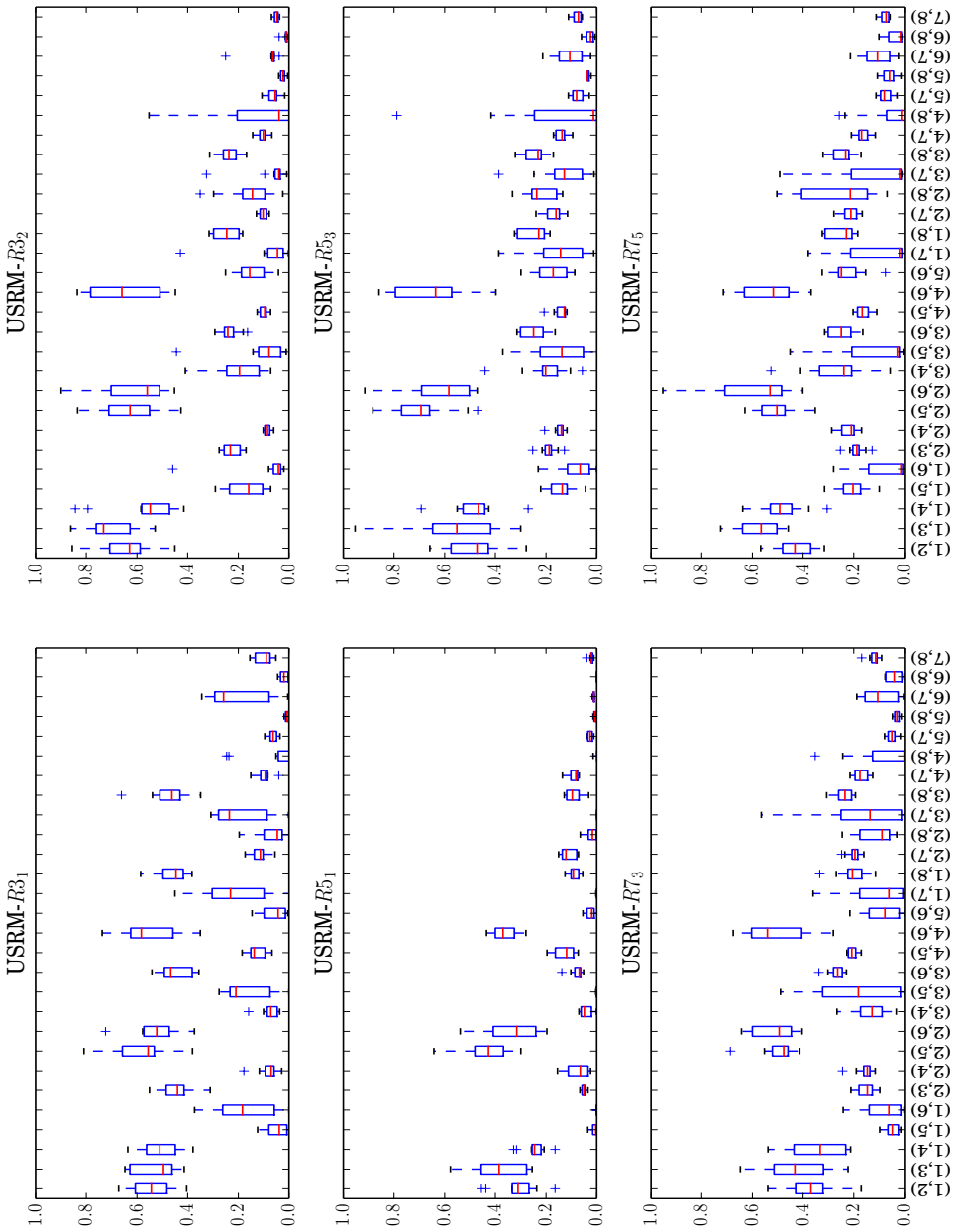

Figure 4.5: The Omega Index of all pairwise multiplex-communities detection algorithms on USRM-R3$_1$, USRM-R3$_2$, USRM-R5$_1$, USRM-R5$_3$, USRM-R7$_3$ and USRM-R7$_5$. The behavior of the algorithms is drastically different from the 2-dimensional cases.

Neighbors are clearly different since the former concerns the connectivity between vertex pairs whereas the latter concerns the connectivity between the neighbors of vertex pairs. This difference is more apparent for higher dimensions. However this disparity does not appear in real-world data (section 4.7.4).

It is particularly interesting that although Projection-Average and Cluster-based Similarity Partition are not similar, they are both relatively similar to CLECC Bridge Detection. Moreover only at higher dimensions Projection-Average is similar to Generalized Canonical Correlations. There is no strong argument to this statistical observation besides that these algorithms follow the general strategy to prefer vertex pairs that are similar locally.

4.7.3 Structured Synthetic Random Multiplex

In the previous section, the USRM suggests that $\{\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3, \mathcal{A}_4\}$ yields similar communities. However Fig. 4.6 shows that the SSRM communities from $\mathcal{A}_4$ (Cluster-based Similarity Partition Algorithm) are distinct from the class of projection algorithms ($\mathcal{A}_1$ to $\mathcal{A}_3$). Thus Cluster-based Similarity Partition Algorithm does provide an alternative perspective for multiplex-communities, and it is not a projection algorithm in disguise.

Furthermore in the previous section, Fig. 4.3 and 4.4 show a high similarity variance between $\mathcal{A}_6$ (CLECC Bridge Detection) with projection algorithms. USRM benchmark (Fig. 4.6) is able to emphasize this observation as the score ranges from NMI = 0 to as high as ≈ 0.8. This highlights the cons of the CLECC Bridge Detection algorithm.

CLECC occasionally yields the components prematurely, where one of the components is a small cluster of vertices or even a single vertex as a community. Therefore it appears that CLECC yields significantly different communities since the rest of the algorithms tend not to return small communities. Hence if we exclude such cases, CLECC is quite similar to the projection algorithms for SSRM.

Finally we will compare the algorithms with the "ground-truth" partitions $\mathcal{P}_1$, $\mathcal{P}_2$ and $\mathcal{P}_3$. Table 4.4 shows that none of the algorithms were able to capture $\mathcal{P}_1$ sufficiently well, which is the partition with high redundancy. Although $\mathcal{A}_3$ (Projection-Neighbors) was proposed to extract high redundancy communities [16], it was $\mathcal{A}_4$ (Cluster-based Similarity Partition Algorithm) that has the best result.

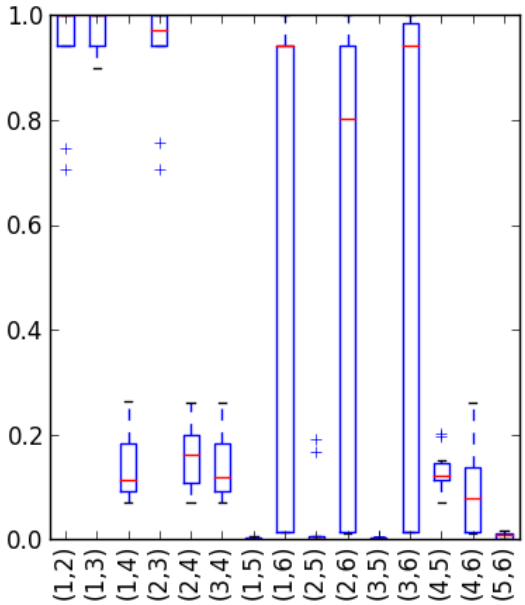

Figure 4.6: The NMI scores of all pairwise non-overlapping multiplex-communities detection algorithms for SSRM benchmarks.

	$\mathcal{P}_1$	$\mathcal{P}_2$	$\mathcal{P}_3$
$\mathcal{A}_1$	0	**0.983**	0
$\mathcal{A}_2$	0.002	**0.94**	0.017
$\mathcal{A}_3$	0	**0.978**	0
$\mathcal{A}_4$	**0.019**	0.14	0.083
$\mathcal{A}_5$	0.004	0.002	**0.158**
$\mathcal{A}_6$	0.006	**0.964**	0.006

Table 4.4: The NMI scores between the algorithms and the different ground-truth partitions. The entries in bold represent the algorithms that are closest to the ground-truth partition.

4.7.4 Real World Multiplex

The results for the European Air Transportation Network are similar to the Youtube Social Network (Fig. 4.7), hence the discussion on either one is sufficient. The general observation is similar to USRM 2, 4 and 5 in Fig. 4.3, where the set of algorithms $\{\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3, \mathcal{A}_4, \mathcal{A}_7\}$ are relatively similar with pairwise NMI score of ≈ 0.55.

In addition, Fig. 4.7 highlights that overlapping-communities detection algorithm $\mathcal{A}_8$ (PARAFAC) is relatively more similar to $\mathcal{A}_5$ (Generalized Canonical Correlations) than the other non-overlapping communities detection algorithms. This observation was less apparent in Fig. 4.3.

Unfortunately $\mathcal{A}_6$ (CLECC Bridge Detection) tends to halt prematurely despite different parameter choices. Hence we did not manage to get any insight for CLECC Bridge Detection in this experiment.

4.7.5 General Observations

The parameters in algorithms $\mathcal{A}_6$, $\mathcal{A}_7$ and $\mathcal{A}_8$ require manual fine-tuning to yield meaningful communities for comparisons. Hence it is not practical to exhaustively test for all configurations for these algorithms. However the analysis does not change in any essential way when the experiments are made for different parameters choices.

From USRM benchmarks and real world multiplexes, $\{\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3, \mathcal{A}_4, \mathcal{A}_7\}$ tend to generate relatively similar partitions. However SSRM benchmark demonstrate that $\mathcal{A}_4$ is able to capture high redundancy communities and performed differently from the class of projection algorithms.

Our experiments with USRM and SSRM benchmarks support that $\mathcal{A}_6$ (CLECC Bridge Detection) is similar to the class of projection algorithms $\{\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3\}$ when the algorithm does not infer small clusters of vertices as communities. Therefore CLECC Bridge Detection is not particularly insightful and not very stable without careful and manual adjustments to the algorithm's parameters.

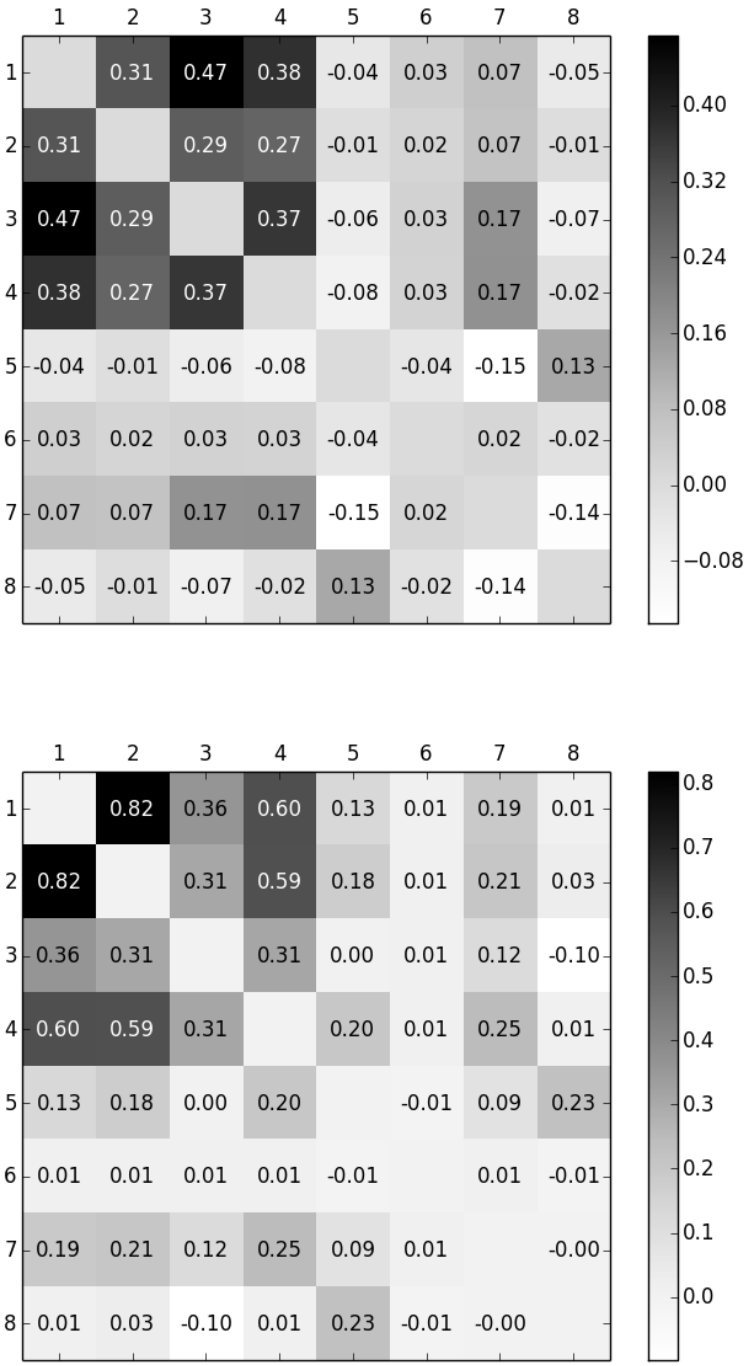

Figure 4.7: The Omega Index heatmap of all pairwise algorithms for the European Air Transportation Network (top) and Youtube Network (bottom).

4.8 Recent Developments

During the writing up and revisions of this thesis, more related research (in preprints) surfaced into the literature. Some are relatively new findings and some are from disciplines that were not considered/explored during our investigations. Regrettably it is hard to include them without major rewriting, simulations and the time to review these preprints. However for completeness the following is an outline of these research.

- Barzinpour *et al.* proposed a spectral approach to communities detection. It is similar to Generalized Canonical Correlations (section 4.4.2) where the multiplex is mapped to a Euclidean Space and communities are found using k-mean clustering. They have also introduced closeness centrality of multiplexes from the communities [12].

- Multiplex communities detection can be presented as a heterogeneous data clustering problem in computer science (database management). Thus using their specialized tools like Relational Bayesian Networks, one can derive the communities [85].

- Hao *et al.* proposed a metric (impact-strength-index) to measure the influence of a monoplex-community in one of the auxiliary-partitions has on the monoplex-communities of the other auxiliary-partitions [77].

- Zhu and Li proposed a projection algorithm. The first step is to quantify the importance of every monoplex by measuring how correlated one monoplex is to the rest of the multiplex. The measure of importance will be used for the weighted average in the next step. Since every monoplex yields a similarity matrix between all pairwise nodes, the projected network is the weighted average of these matrices [168].

- *MutuRank* by Wu *et al.* is also based on the strategy of projection. It uses both the probability distributions that a vertex chooses its neighbors and relationship to form a distribution on the frequency of the relationships. This frequency distribution is then used to project the multiplex in a linear way [165].

- Infomap is a monoplex-communities detection algorithm that is based on the compressibility of a random walk. Domenico *et al.* extended the idea by factoring the probability of swapping relationship into the nodes' transition probability [54].

- Bródka and Grecki proposed a benchmark multiplex based on a well known network benchmark — LFR Benchmark. [29].

- Nicosia *et al.* described two multiplex constructions that is based on preferential attachment. Every new vertex that is introduced iteratively into the multiplex is given a fixed number of edges in every relationship. However at the different monoplex, the new edges will connect to the rest of the vertices at a different time [131]. Alternatively different relationships can have different preferential attachment behaviors while maintaining certain correlations [132].

4.9 Summary

This chapter summarizes the different multiplex-communities in the literature and compares the algorithms that are used to find these structures. The emphasis of this chapter is to distinguish the different multiplex-communities and show how these solutions deviate when we have less-than-ideal situations.

This is due to the limits of abstraction to model the relational structures like communities. The multifaceted definitions of multiplex-communities are based on anecdotal observations of real world data which can be hard to translate as a general framework. Ergo the methodology of a case study cannot be easily transfer to other situations.

For example let $\mathcal{M}_1$ be a social-multiplex on $\{$*colleagues, family*$\}$ and $\mathcal{M}_2$ be a citation-multiplex on $\{$*references, common keywords*$\}$. If $\mathcal{A}_1$ and $\mathcal{A}_2$ are the algorithms that were used to study $\mathcal{M}_1$ and $\mathcal{M}_2$ respectively, then is there a systematic and objective way to determine which of the two algorithms is better for a transportation-multiplex on $\{$*buses,trains*$\}$? Moreover is it even possible to bootstrap the fact that $\mathcal{A}_1$ is the right algorithm for $\mathcal{M}_1$?

Given the subjective nature of relationships, it is difficult to suggest a convincing framework to address the above issues. The appeal of a multiplex is that it preserves the fine relational properties of a system, but the trade-off requires additional assumptions on the model. Moreover since the multiplex-communities detection algorithms are very sensitive, a slight error on any of the assumptions can easily lead to the wrong results.

Chapter 5

The Network Science of Interval Graphs

There could be many reasons to why interval graph research lost traction. A probable cause is that the computational challenges for interval graph algorithms (e.g. boxicity) limits its application and hence losing the support from the natural sciences. Nevertheless the mathematics of interval graphs are thoroughly researched for many years and hence there is very little that this thesis can contribute to the study.

However given the obscurity of interval graph research, there is hardly any modern perspectives to the problems and its *applications*. Thus in the first part of this chapter, we look at a heuristic strategy to compute a graph's boxicity by modularizing the problem with communities detection algorithms. There are three reasons to do so: 1) It helps to reduce the computational hard problem into simpler problems. 2) It gives us multiple perspectives of complexity — the local complexity of the communities and the global complexity of the general topology. 3) More importantly it addresses the practical issues in experimental science where the noise in the data affects the boxicity of the network.

The second part of this chapter is a framework with interval graphs to model the discontinuity of information propagation. The hyperbox representation of a graph is a deterministic model that is able to simulate information flow between non-adjacent vertices. This is different from current models where the discontinuity effect is often obscured by random processes. This is still in the preliminary phase and is left for future work.

The research presented in this chapter is published in [112].

5.1 "Approximating" Boxicity Using Communities Detection

The boxicity of a network can be used as a measure of complexity. It determines the minimum number of attributes to measure the vertices such that an adjacent vertex pair implies that their attributes overlap. Given that an adjacent vertex pair in a multiplex is connected by relationships, it suggests that the vertices must have common (overlapping) attributes such that the relationship can be established. Therefore we posit that **the number of relationships in a multiplex $\mathcal{M}$ should be less than the boxicity of its projection.**

The rational is that if we assume linear structures like interval graph as the simplest type of relationship, then the complexity will unnecessarily increase when we introduce more relationships to the multiplex. The assumption is that it takes at least one metric to measure a relationship. Hence if d metrics is sufficient to describe a network, then a multiplex with more than d relationships implies that there are some dependency among the relationships.

For example let the relationships in a multiplex be $\{R_a, R_b, R_c\}$ and the metrics to measure these relationships be $\{a, b, c\}$ respectively. Suppose the boxicity of the multiplex's projection is two. This implies that the metrics are not independent, e.g. let a be *temperature* and b is *energy*, thus R_a is redundant as temperature is a type of energy. Alternatively there could be a better relationship to describe the system, e.g. let a be *temperature* and b be *the distance from the sea*, hence a simpler metric is *humidity*. The difficulty is to establish the identity of the relationship that is described by the metrics.

This simplification of multiplexes is the same motivation as De Domenico *et al.* in [55]. The idea is to simplify the complexity of a multiplex by reducing the number of relationships, while maximizing the information. However the applicability of boxicity for real world systems will remain limited since to determine a network's boxicity is a NP-complete problem.

Although a network's boxicity is bounded by a function on the graph's properties (Table 2.1), the bound is generally not tight and thus it is still not very meaningful. In addition the boxicity of a network is an unstable and non-monotonous function where it fluctuates unpredictability when new edges/nodes are added to the network. Hence the intolerance to experimental errors further challenges the applicability of boxicity. This section resolves the above issues using communities detection to modularize the problem.

5.1.1 Minimum Boxicity of Network from its Communities

We propose that communities detection is a key strategy to approximate the boxicity of a network. It is similar to optimizing the Hamiltonian Walk problem by simplifying a network into modular structures [35]. Clearly the boxicity of a community is a simpler problem since it is a smaller graph. However more importantly we can bound the boxicity of the entire network using the boxicity of the communities [140]:

Lemma 5.1.1 *$Boxicity(G) \geq max_{g \in C} Boxicity(g)$, where C is the set of all the communities in graph G.*

For instance there are two communities in the Zachary Karate Club Network with 17 vertices in community A and 16 vertices in community B (Top diagram in Fig. 5.1). The network is not an interval graph as vertices $\{24, 25, 26, 28\}$ is not chordal (theorem 2.1.2). Since the communities are small, we are able to deduce (via exhaustive search described in section 2.1.2) that community A and B has boxicity $= 2$ and > 2 (no solution found) respectively.

Since community B is a planar graph, its boxicity ≤ 3 [162] or more specifically $= 3$. Thus the boxicity of the Zachary Karate Club Network ≥ 3 (lemma 5.1.1). The bottom diagram in Fig. 5.1 shows one of the possible hyperbox representations of the communities. Since we have to eventually combine these partial solutions, it appears to be very helpful to constrain the partial solutions such that the vertices of a community that connects to the other communities have to be aligned along the boundaries of the hyperboxes.

Fig. 5.2 slightly rearranged the hyperboxes in Fig. 5.1 such that the boxes at the boundaries the communities can be easily combined. From the figure, we can conclude that it does not take more than 3-dimensions to combine the communities. Hence the boxicity of Zachary Karate Club Network is 3.

5.1.2 Boxicity of the Communities' Interaction Network

The communities also allows us to look at the broad overview to how information flows from one community to another. That is the complexity (boxicity) of the modular structures is important to understand the information propagation of a network (section 5.2).

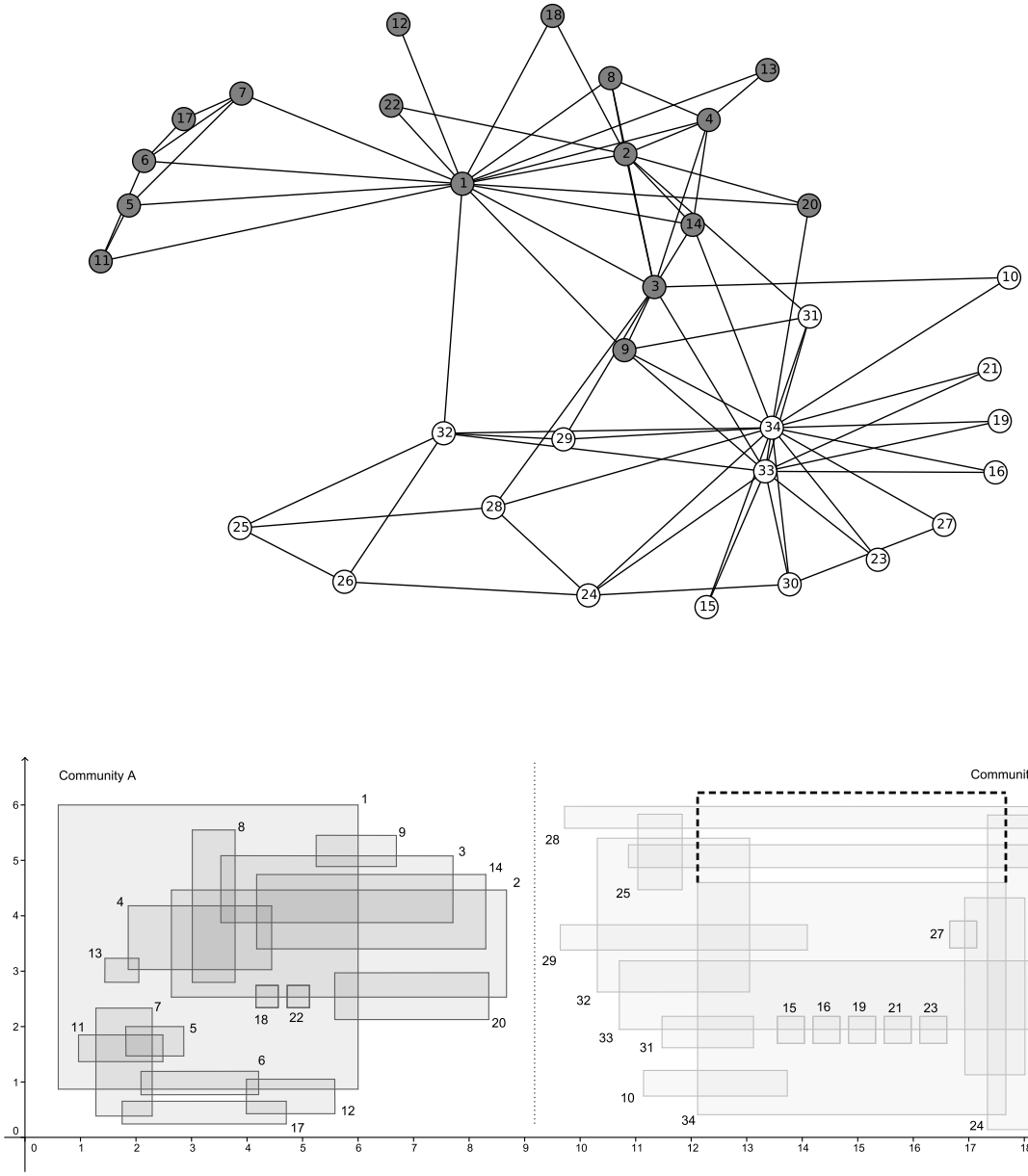

Figure 5.1: (Top) The Zachary Karate Club Network where communities A and B are the shaded and non-shaded nodes respectively. (Bottom) The hyperbox representation of community A and B. The dashed line represents Community B in 2-dimension (boxicity $= 2$) if vertex 34 is adjacent to vertex 26. Since it is not, hence we need the third dimension such that the box 34 can overlap box 28 while “bridging over” (bypass) box 26. The boxes are aligned in a way such that vertices that connects to the other communities are near the center. For example the vertices in community A have to route via vertices $\{1, 2, 3, 9, 14, 20\}$ to get to community B. Similarly the vertices in community B have to route via vertices $\{10, 28, 29, 31, 32, 33, 34\}$ to reach community A. Since we divide the network into two smaller communities, this constrain is more sensible when we try to “join” the communities’ hyperboxes.

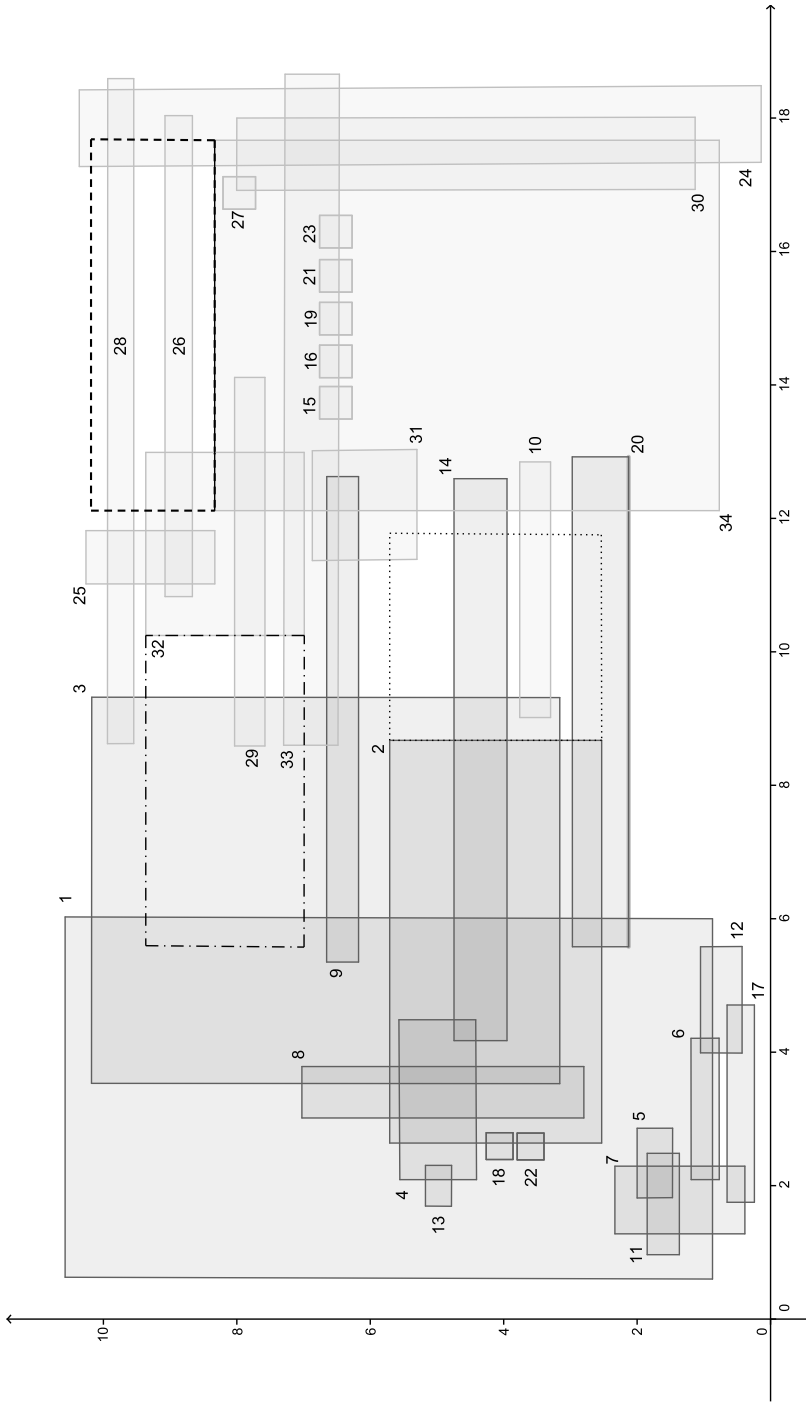

Figure 5.2: The hyperbox representation of the Zachary Karate Club Network. The boxes have to be transformed such that the connection between communities A and B is complete. The dashed line refers to the extension of box 34 to overlap box 28 such that it bypass box 26. The dotted line refers to the extension of box 2 to overlap box 31 such that it bypass box 10. The dot-dashed line refers to the extension of box 32 to overlap box 1 such that it bypass box 3.

This is done by coarsening the network G with a new network H where the vertices in H represents the communities of G, and the vertices in H are adjacent if and only if their corresponding communities in G are connected. Again, since H is a smaller network than G, the computation of the boxicity of H is easier. This process can be repeated on H until we get the desired granularity.

In our previous example, the Zachary Karate Club Network, there are only two communities and they are connected. Hence the coarse network is just a complete graph on two vertices, i.e. an interval graph with boxicity $= 1$. This implies that the information flow has low complexity where there is a linear flow from one community to another.

5.1.3 Boxicity with Experimental Noise

The conclusion from the previous example is trivial since there are only two communities. However it is interesting and important to note that the conclusion remains the same when we remove or add (a small number of) edges from/to the network. These modifications can represent the noise in the experiments and hence more relevant for scientific applications.

Quasi-Interval Graph, Q is a graph with boxicity > 1 that can be expressed as an interval graph by adding or subtracting some edges as experimental errors from Q. It is useful for systems where there are strong qualitative evidences that they have linear structure [40, 122, 158]. This can be done by finding the minimum number of edges to 1) add to Q [88, 89, 127], 2) remove from Q [72] or 3) a mixture of both types of errors [116]. For example community B in the Zachary Karate Club Network will have boxicity $= 2$ (instead of 3) if vertices 26 and 34 are adjacent (Fig. 5.1).

However it is still a hard problem to minimize the number of modification over the entire network such that its boxicity is also minimized. Thus is more intuitive and easier to understand the general dynamics of a system with the coarsen network than the precise boxicity of the network.

5.2 Information Propagation of Interval Graphs (Future Work)

Information propagation is the behavior in which a property on the vertices is spread across the graph. In the *infection model*, a vertex passes the property to its neighbors probabilistically at each iteration. This models the behavior of a virus epidemic where there is a probability for an entity to catch the virus from its neighbor [7, 80].

Alternatively a vertex adopts the property under the influence of its neighbors when the ratio of its neighbors with the property exceeds a threshold. This is the *influence model* and it is used to describe the nature of social trends like product recommendations [21, 73, 75]. In general terms, vertices with the information (e.g. infection) are known as *active* vertices, and if otherwise they are known as *inactive* vertices.

A common notion with these models is that information flow along the edges of the network. However it is not possible to consider all the relationships in the system to map the full topology of the network. Thus it is possible for information to flow between non-adjacent vertices. This discontinuous flow of information is often assumed to be the actions of some confounding variables in the system and is simulated by passing the information probabilistically to a random non-adjacent vertex [124, 144]. This section as future work, proposes the hyper-boxes representation of a graph as a deterministic linear framework to model the discontinuous flow of information.

5.2.1 Outline

The linear fine structures of a graph G is the set of m interval graphs $\{I^1(V, E_1), \ldots, I^m(V, E_m)\}$ as hyper-boxes where $G(V, E) = (V, E_1 \cap \ldots \cap E_m)$. The set of edges from the intervals graphs that are not in G, i.e. $E^c = (E_1 \cup \ldots \cup E_m) \setminus E$, are the confounding edges unobserved from the graph G. Thus when information flow through the edges in E^c, it will appear from G that there is a discontinuous flow of information (Fig. 5.3).

For example in a marine food web, a predator feeds on two environment niches/metrics — the size of the prey and the depth of the ocean where the predator hunts. Hypothetically suppose there is toxin deposits in the ecology and via bioaccumulation the toxin level of a fish is proportional to its size. Thus the spread of the toxin will appear discontinuous from the food web since the feeding patterns in deep water is different from the surface.

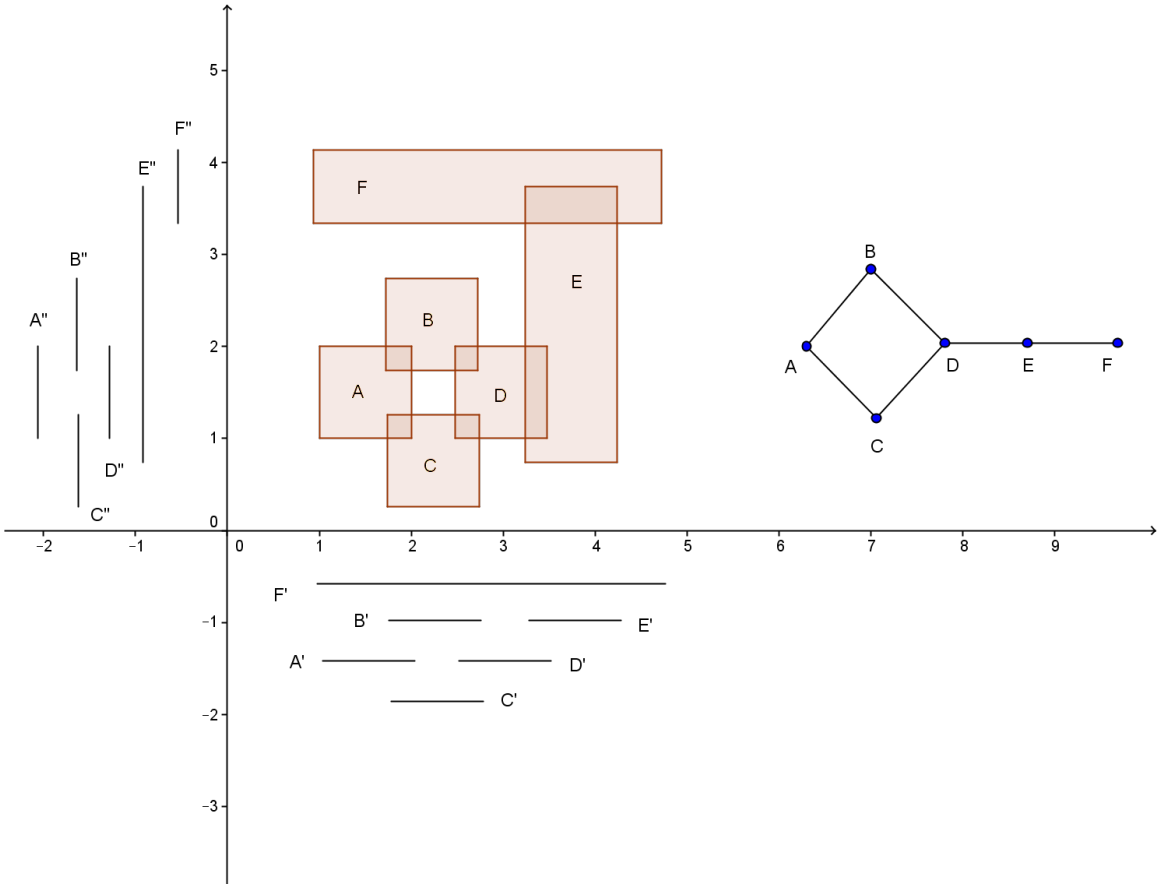

Figure 5.3: (Note: the same diagram in Fig. 2.1.2) The box representation (middle) of a graph (right) with boxicity 2, where the label of the boxes correspond to the vertices of the graph. The box (graph) is the intersection of the bottom and left intervals, where box A is enclosed by intervals A' and A''. Suppose interval A' (vertex A) is active and infects adjacent interval F'. Although A' and F' are adjacent, their respective boxes (vertex) are not adjacent, i.e. A is not adjacent to F. Hence from the perspective of the graph, there is a discontinuous flow of infection between non-adjacent vertices.

This framework does not obscure the context of the propagation's dynamics by random process, i.e. the flow of information is well defined either via the edges of E or E^c. However the trade-off is the computational intractability to derive the hyper-box representation from a given graph. Thus in the experiments, the random interval graphs are first constructed and then their intersection forms the observable graph G.

For simplicity we will only consider the case where the G is connected. This condition can be met by varying the radius of evolutionary interval graphs (section 3.6). In addition we can parameterize the radius so that we can observe different rates of discontinuity.

5.2.2 Propagation Models

Given a connected graph $G = J_r^1 \cap \ldots \cap J_r^m$, the propagation dynamics are applied to one of the evolutionary interval graphs J_r^k. For example in the infection model, the active vertices in J_r^k infects their neighbors with fixed probability. Since the neighbors in J_r^k are not necessarily adjacent in G, the discontinuity of information flow can be observed from the perspective of G, which means that information flow is disrupted in G.

Infection Model

The framework of a typical infection model (SIR: susceptible-infectious-recovered) is the process where active vertices can transmit the infection to inactive vertices with a fixed probability per unit time. Concurrently active vertices can recover at a constant rate. The ratio between the infection rate and the recovery rate determines the spread of the infection (epidemic) across the network [46].

However the rate of recovery is not required to observe the discontinuous flow of information. This simplification is analogous to the spread of news or gossips across social networks via word of mouth [152]. The rate of infection follows the assumption that an inactive vertex v is more susceptible to be infected if most of its neighbors are active. Thus

$$Pr(v \text{ will be infected}) = \frac{\text{No. of active neighbors}}{\text{No. of neighbors}}, \tag{5.1}$$

and discontinuity is defined when a vertex is infected with zero probability.

Influence Model

In the influence process, an inactive vertex in a network becomes active if a sufficient ratio, τ of its adjacent vertices are active. It is similar to the behavior of fashion trends in social networks where "non-adopters" (inactive) vertices follows the style under the influence of their peers. Hence much more active vertices are required to influence a high degree vertex than a vertex with fewer neighbors. Therefore it is possible to reach an equilibrium when information no longer spread across the network, where there are insufficient active vertices to influence remaining inactive high degree vertices [93].

In the experiments, a vertex will be active if at least half of the neighbors have to be active, i.e. $\tau = 0.5$. From the perspective of the graph G, if an inactive vertex becomes active when half of its neighbors are inactive, then this situation is defined as an instance of discontinuity in the information flow. Conversely, discontinuity is also defined if a vertex remains inactive even when it is above the threshold.

5.2.3 Experimental Results

Fig. 5.4 shows a proof of concept that it is possible to simulate any rate of discontinuity with evolutionary interval graphs. We define the rate of discontinuity $= 1$ when the graph is disconnected so that the plot fits a Sigmoid-like function. Furthermore as r increases, the graph becomes denser and every vertex has an edge connecting to most of the other vertices. Thus the effects of discontinuity is less apparent, although the observed frequency is different from the expectation probability of Eq. 5.1.

In real-world systems, interval graphs do not necessarily belong to the ensemble of evolutionary interval graphs. This experiment is simply to show that the framework is able to model the discontinuity in network propagation. The advantage of a deterministic model is that the simulation behaviors are repeatable once the general direction of the flow is fixed.

The emphasis is that this framework is an *alternative* and not a *replacement* for existing models. Given that intervals graphs have meaningful contexts in complex systems like Ecology and Bioinformatics, future work is to find valid applications in the broader scope of Network Science such that this framework aptly models the system. For example time dependent systems like the EEG or fMRI time series of brain networks.

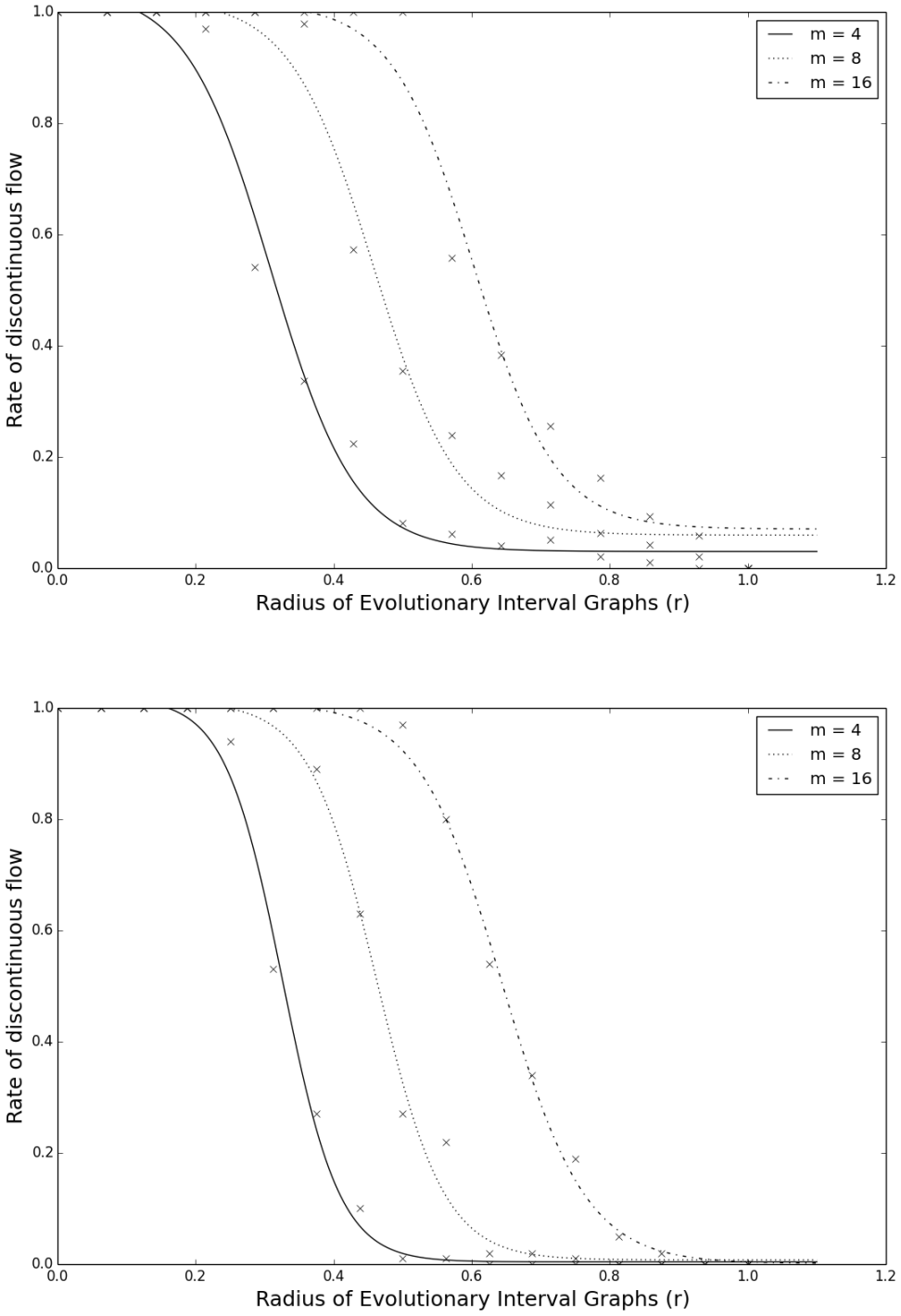

Figure 5.4: The rate of discontinuity observed in $G = J_r^1 \cap \ldots \cap J_r^m$ when the infection (top plot) dynamics is applied to a random interval graph J_r^k. Similarly the bottom plot shows the rate of discontinuity when the influence dynamics is applied to a random evolutionary interval graph.

5.3 Summary

Interval graphs are well-studied in ecology to understand the stability of complex systems. However to relate this to the broader applications in Network Science has yet to be attempted. For instance the simulations show that as a proof of concept, interval graphs are viable linear fine structures to model the real world characteristic of discontinuous information propagation. The advantage of this framework is that the intuitions of the dynamics are not obscure by random processes.

In addition we show that the methodologies in Network Science can help in the computational challenges of interval graph — modularize a network such that the computation problem of boxicity can be simplified. Furthermore the communities allow us to focus on the complexity of the general network topology rather than the details within the communities which are prone to experimental errors.

Given the growing interests for multi-relational networks, interval graphs provide an alternative model for scientists in their research. In addition the modern methodologies and tools from Network Science can address the computational challenges of interval graphs problems.

Chapter 6

Disparate Literature

Section 2.2.1 briefly mentioned the disparate nature of this study, which posed a huge challenge to sieve through the synonymous names in the literature. Moreover these names are interchangeably used to describe other generalized graph models that are vaguely similar but mathematically different.

For example *multilayer networks* is one of the synonymous names in the literature, where the name can also refer to a set of graphs (as layers of graphs) that is *not necessarily* on the same vertex set with an *interlayer edge set* to connect the graphs in the set. This is used to model a set of somewhat independent systems that has subtle dependency between them, e.g. a power-grid network failure will disrupt the power required for a communication network [145].

Therefore for the completeness of this thesis, this chapter introduces these mathematical objects as a broad overview of the disparate layered-network research and then applied *Mark-and-Recapture* from population biology to estimate the number of publications that were overlooked in the current literature reviews.

Finally we will look into further applications of Mark-and-Recapture as a stopping rule for the results from search engines. This also allows us to measure the similarity of truncated-rankings using the idea from the Least Square Error.

The research presented in this chapter is published in [110].

6.1 Formal Definitions of Other Generalized Models

Definition 6.1.1 *The most general form of a layered network is a finite set of m graphs as layers $\mathcal{G} = \{G^1(V_1, E_1), \ldots, G^m(V_m, E_m)\}$ with*

$$\mathcal{C} = \{\bar{E}_{ab} \subseteq V_a \times V_b; a, b \in \{1, \ldots, m\}, a \neq b\} \tag{6.1}$$

as the set of crossed layers *elements to connect the vertices of one layer to the vertices of another, i.e.* interlayer *connections [23, 95].*

The most general form of a layered network is a set of graphs as layers and the graphs are connected by interlayer edges (Def. 6.1.1). For instance a multiplex is the variant where the graphs are on the same vertex set V and do not have interlayer connections ($\mathcal{C} = \emptyset$). Similarly the other variants can be defined as follows:

1. An ordered (with respect to time) set of the graphs in a multiplex is known as a *temporal network* where it models the change of the connectivity in a system.

2. A strikingly similar variant to multiplex is the case where the graphs are on the same *copy* of vertex set V, i.e. $V_1 = \ldots = V_m = V$, and the graphs are connected by interlayer edges $\bar{E}_{ab} = \{(v, v); v \in V\}$.

3. A *multilevel network* is a set of graphs $\{G^1(V_1, E_1), \ldots, G^m(V_m, E_m)\}$ on vertex V where the vertex sets V_i are subsets of the copies of V such that $V = \cup_1^m V_i$. The interlayer edges are $\bar{E}_{ab} = \{(v, v); v \in V; v \in V_a \cap V_b\}$.

4. A *hypergraph* is a generalization of a graph in which a hyperedge $\bar{e}_i$ connects multiple vertices. It is a type of *multilevel network* $\{G^1(V_1, E_1), \ldots, G^m(V_m, E_m)\}$, where each edge $\bar{e}_i$ forms a complete graph G^i on all the vertices connected by $\bar{e}_i$.

It is hard to determine which model best describes a system since under certain conditions the models can be expressed interchangeably with little or no loss of information, e.g. a hypergraph as a multilevel network. Thus literature reviews like [23, 95] present these variants of layered networks as a unified framework.

6.2 Population Estimation

It is highly probable that the bibliography of the literature reviews on layered networks are incomplete. Just like population biology, it is not possible to capture all the animals to determine the population of an animal species. Instead a *mark-and-recapture* methodology can be used to approximate the population by multiple samplings of the species and discount for the number of instances that were caught previously.

6.2.1 Mark-And-Recapture

Animals are captured and marked before releasing them back in the wild. After enough time has passed to allow a complete mixing, the population is sampled for the second time. In the second sample, the ratio of marked animals (from the first capture) to the number of captured animals is approximately the ratio of captured animals in the first sample to the total population, hence by Peterson method [156]:

$$\text{Total population} \approx \frac{N_1 N_2}{R}, \tag{6.2}$$

with standard deviation

$$\sigma = \sqrt{\frac{(N_1+1)(N_2+1)(N_1-R)(N_2-R)}{(R+1)^2(R+2)}}, \tag{6.3}$$

where N_1, N_2 are the number of captures in the 1^{st} and 2^{nd} sample respectively, and R is the number of marked animals (individuals that were captured in both samplings). Furthermore Mark-And-Recapture can be extended to multiple captures by a weighted average of Eq. 6.2 known as Schnabel Index [151]:

$$\text{Total population} \approx \frac{\sum_{i=1}^{m} N_i M_i}{\sum_{i=1}^{m} R_i}, \tag{6.4}$$

where N_i is the number of captures in the i^{th} sample, M_i is the total number of marked animals in the population before the i^{th} sample, and R_i is the number of marked captures in the i^{th} sample.

6.2.2 Assumptions in the Estimation

To apply the same methods to scientific literature citation analysis, the assumptions have to be parallel to biology population. The mixing period for population biology has to be long enough such that the second sampling is independent from the first, yet short enough to minimize the effects of population changes or the death of the tagged animals, i.e. a closed population. Since the literature review of [23, 95] were independent efforts and were completed approximately at the same time, the sampling is well mixed and it is reasonable to assume that the number of relevant literature is fixed (i.e. closed).

However the probability that a paper is found and referenced is not equal [14]. There are many factors that affects the visibility of a publication, e.g. quality of research, keywords, publication date, authors, etc. This is a common violation of assumption in wildlife as some individuals have a higher tendency to be captured again, i.e. "trap-happy". In such cases, the result will be the lower bound figure to the true population size.

6.2.3 Methodology

There are two sources that throughly review the literature on multilayer networks — *The structure and dynamics of multilayer networks* by Boccaletti *et al.* [23] and *Multilayer Networks* by Kivelä *et al.* [95]. The bibliographies were filtered so that only the publications on layered networks were analyzed. The population is estimated using Eq. 6.2 since the sample size is sufficiently large, i.e. Chapman estimator is similar to Peterson method.

It is important to manually curate the bibliographies as the references might be mislabeled or not up to date. E.g. Kivelä *et al.* referenced the arXiv link to the paper *Metrics for the analysis of multiplex networks (2013)* whereas Boccaletti *et al.* updated the publication in Physical Review E. There are also cases where the authors' names are wrong, e.g. Saramaäki is missing in Physics Reports article *Temporal networks* (2012).

In the event that there are more literature surveys on layered networks, Schnabel index (Eq. 6.4) can be applied to increase the accuracy. Although it is important that the subsequent surveys built upon of the reviews by Kivelä *et. al* and Boccaletti *et. al*. The mechanics of Mark-And-Recapture is a sequence of capturing and marking the animals, hence the number of marked animals is dependent on the prior captures.

6.3 Empirical Results

As a proof of concept, we have to apply this methodology on a different context first. This is to increase our confidence if there are other independent experimental support. Hence we will first apply this idea on the literature on the community of graph (not multiplex).

6.3.1 Literature on the Communities of Graphs

There are several reviews on the community of graphs — Newman 2004 [129], Fortunato and Castellano 2007 [65], Schaeffer 2007 [148], Porter *et al.* 2009 [137], and Fortunato 2010 [64]. Although it is tempting to apply Schnabel Index to sample the body of literature repeatedly, it violates many assumptions of the estimator.

The first violation is that these surveys are not independent sampling of the literature as almost all of them cited the earlier reviews. Secondly the population in question is not closed as there are many research on communities detection since year 2004. There are only 44 references in Newman 2004 review versus the 457 references by Fortunato 2010. Thus the results will be meaningless even if the numbers support the methodology.

Therefore to minimize the violation of the assumptions, the reviews must be published approximately the same year and the latter did not cite the earlier review. Hence Schaeffer 2007 will be the first sample and the review by Fortunato and Castellano 2007 will be the second. Finally the result will be compared against the bibliography of the review by Fortunato 2010 to gauge the accuracy of this methodology.

Out of the 249 references in Schaeffer 2007, only 43 articles are directly relevant to communities detection. Most of the excluded references are on graph cutting from graph theory or clustering algorithms from machine learning, since they do not connote the idea of modularity of communities in the articles. Similarly only 55 articles were chosen from 97 references in the review by Fortunato and Castellano 2007.

Finally there are 20 citations that were listed in both reviews, thus Eq. 6.2 and Eq. 6.3 suggest that there are $\approx 118 \pm 14$ publications on graph communities by 2007. In comparison, there are 112 articles before 2008 on graph communities in the bibliography of Fortunato 2010. This is a close estimate and supports the framework to study the disparate literature on generalized graphs in the next section.

6.3.2 Literature on Generalized Graphs

There are 376 entries in the bibliography in the review by Kivelä *et al.* and among them 233 papers are directly related to layered networks. The review by Boccaletti *et al.* is more comprehensive in describing the background of the problems, hence it has 520 articles in the bibliography but only 214 of them are on the layered networks.

Both surveys did a thorough review on the formalism and the relationships amongst the different generalized graphs. However Kivelä *et al.* has more coverage on the multiplex communities whereas Boccaletti *et al.* is more comprehensive on the synchronization problem and provides more examples in real-world applications.

Therefore approximately 60% of the bibliographies overlap, specifically 143 common relevant articles. Thus from Eq. 6.2, the lower bound to the number of articles on layered networks is $233 \cdot 214/143 = 348.68 \approx 350$ with standard deviation of $\sigma \approx 10$. The total number of articles from both review is $233 + 214 - 143 = 304$, hence there is potentially at least $350 - 304 = 46$ relevant literature that were overlooked in the process. Although there are much more "missing-relevant" articles in Graph Theory (i.e. interval graphs), it is reasonable that most of them are not directly relevant to the applications of multiplexes.

6.4 Application in Bibliographic Search

Since literature reviews are well curated, the estimate from Mark-And-Recapture suggest the size of the body of literature on a given topic. It gives new researchers a level of confidence in their preliminary investigations.

However the conditions for this methodology are hard to meet (section 6.2.2) for most research topics. Furthermore it begs the question, *is the bibliography of the literature reviews complete?*. Since academic search engines are the basic sources of information for researchers, it would be interesting to apply Mark-And-Recapture to compare the results from the different search engines.

The preliminary process of a research is the task of searching and *re-searching* the relevant publications to have an overview on the topic. There is no optimal stopping rule to determine if one has collected sufficient relevant articles, and prolonged search will tend to

have diminishing returns. This is a foremost challenge for any researchers and one of the reasons for peer reviewing publications (i.e. to avoid duplicated research).

For instance although there is a huge body of research in medical science, there is an urgency to provide the proper medical care. Thus the time spent on research has to be optimized. However the citation network for clinical trials is disconnected, which reflects the possibility that the "different camps" of clinical researchers use different research tools. Hence many of them are unaware of the relevant literature from the other camps [142].

Thus Mark-And-Recapture methodology was proposed as a stopping rule for medical research [25, 90, 91, 102, 157]. For example the empirical evaluation on osteoporosis disease management publications estimates approximately 592 articles are missing from the four main bibliographic databases for medical science — MEDLINE, EMBASE, CINAHL, and EBM Reviews [90].

6.4.1 Comparing Search Engines

Experiment Methodology

Unlike in the medical field, many keywords in science have multiple meanings in different contexts, for example the word *graph* can be defined as a plot of a function or an abstract mathematical object (i.e. network). Hence there can be many unrelated results and causes the search engine to return millions of articles.

One way to sieve through the articles is to accept the top "relevant" articles suggested by the search engine until there is no new significant information is gained [32]. However there is no measure of *information gain* and we depend mainly on our subjective gut feelings. This section shows that Mark-And-Recapture on academic search engines (e.g. Google Scholar, Microsoft Academic Search, and Web of Science) can be used to quantify the "information gain".

The web-crawler/database of these search engines are the "traps" for the entire body of literature, and the ordering of the results is a reflection of the (search engine) algorithms' unique perspectives of the keywords. Suppose the top n^{th} results of two search engines, E_1 and E_2, have R number of common articles. Eq. 6.2 suggests that there is at least $T = n^2/R$ publications on this topic. To avoid the division by zero, we initialized $R = 1$.

If we assumed that one stops at the n^{th} entry of E_1 and E_2, then the coverage of the body of literature is at most $C = (2n - R)/T$. Therefore the rate of change of C with respect to n estimates the information gained during the time spent with the search engines. A low rate of change implies low information gain and quantifies a stop to the search.

For simplicity, this thesis only compares two search engines at a time where each of them is independent sampling over the body of literature. The ordering of the results is sorted by "relevance" which is ranked by the different search engine algorithms.

Lastly only the top 500 results from each search engine are collected in the experiments since Web of Science limits that number of articles to be exported at each time. Moreover if the sampling is too large it will trigger Google Scholar to temporarily ban users from accessing its database. The software used to extract from Google Scholar and Microsoft Academic Search is *Publish or Perish* [79].

Formalism

Some papers are published in multiple sources, e.g. arXiv and peer-review journals and it will cause the search engines to return the same paper as distinct publications. Since there is no information gain for repeated articles, we have to adjust our equations.

The coverage of a literature is a time series where the n^{th} unit of time refers to the n^{th} article of the search engines. Let $N_{i,n}$ be the number of *unique* articles returned by search engine E_i at time n. Then the estimated total number of publications on this topic is

$$T = N_{1,n}N_{2,n}/R, \tag{6.5}$$

where R is the number of articles found in both search engine. Hence the coverage,

$$C = (N_{1,n} + N_{2,n} - R)/T. \tag{6.6}$$

In most cases $N_{1,n} = N_{2,n} \approx n$. If $R \to$ constant, then $\lim_{n\to\infty} C \approx 1/n \to 0$. This implies there is a diminishing return to the information gain when one continues the search. From another perspective the total number of publications $\lim_{n\to\infty} T \approx n^2 \to \infty$. This means that the given keyword is so imprecise that the results from the different search

engines diverge as there is almost no common articles between the search engines.

If $R \approx n$, then it implies that E_1 and E_2 are so similar and it is analogous to using one search engine. In such case we are back to the original situation where there is no quantified method to stop the search. Fortunately R generally does not grow in a fixed manner for the entire time series. There are instances when the derivative of C is zero and it usually infers an optimal stopping rule.

After a local maximum of C, the information gain is negative as the search engines' perspectives of the keyword begin to diverge. Hence the reason to stop is that the subsequent articles are less relevant from the perspective of the other search engines.

After a local minimum of C, the stopping rule is slightly counter-intuitive. Although the coverage increases, R increases rapidly too. This means that the subsequent articles are already returned in (much) earlier results, and hence no information gain.

Types of Stopping Rule

Type I (Convergence to Zero)

The quality of a search depends on how specific the keywords are, for example many disciplines like physics, chemistry and engineering have subfields that research on improving rechargeable batteries. Hence the results from the different search engines are drastically different with keywords like "rechargeable batteries" (Fig. 6.1).

Therefore if a keyword has similar figure, then it suggests that one should refine the keyword to be more specific. The keyword is either too ambiguous like "Phase Transition" and "Communities Detection", or the topic is studied in many branches of science like "Genetic Algorithm" and "Ising Model". In such cases, there is no good stopping rule.

Type II (Local Max and Min)

When T grows quadratically, it means that the search results are drastically different. This usually implies that the choice of keywords is bad and one should discard the search results. However it is not true in general, e.g. consider the keyword "Kauffman Model" in Fig. 6.2.

The local minimum of C (for dotted and dashed line) is approximately at $n = 20$ where T appears to be linear in log-scale (i.e. polynomial growth). The rapid increase of coverage plateaued approximately at $n = 50$ is the effect that the subsequent articles after $n = 20$ in

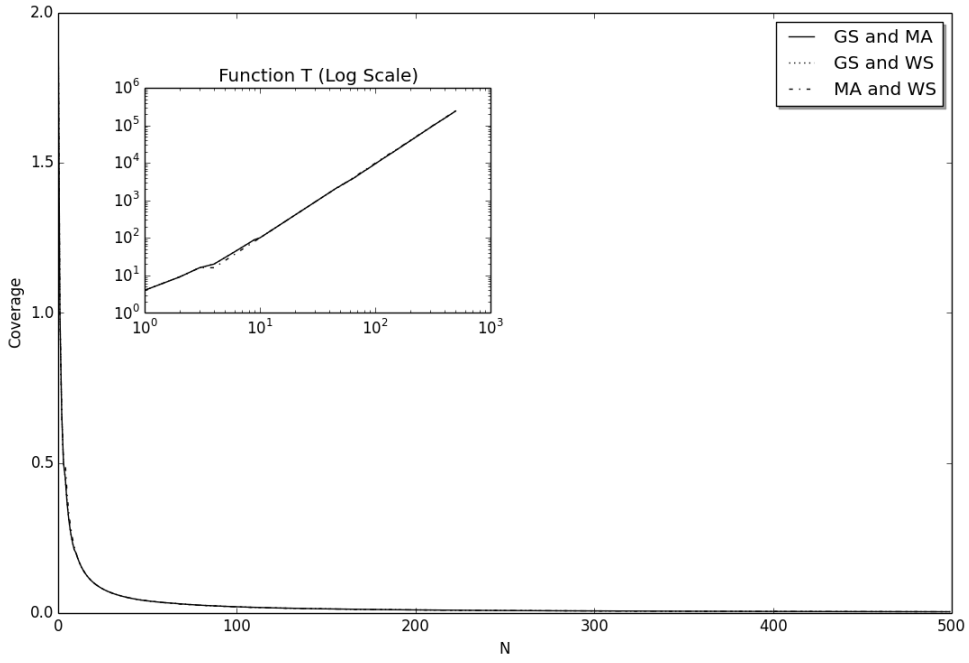

Figure 6.1: **Keyword: Rechargeable Batteries**. GS, MA and WS are abbreviations for Google Scholar, Microsoft Academic Search and Web of Science respectively. T is quadratic for all pairwise comparisons (R grows so slowly that it is almost constant), hence linear in the inserted figure. This implies that a search on general keywords like "rechargeable batteries" are unfocused and can be found in many different fields of research.

one of the search engines were already listed in the search result of the other search engine. Thus there is little information gain and it is a reasonable to stop at $n = 20$.

The local maximum of C plateaued until $n \approx 70$, where it is an alternative stopping point for the search. It is an indicator that the search engines' suggestions begin to deviate and hence subsequent articles are less relevant.

Keywords with graphs that are similar to Fig. 6.2 are unfortunately not very common. Out of the 50 keywords selected for our experiments, only the graphs of "Kauffman model" and "Tangled Nature Model" have both local minimum and maximum.

Type III (Local Min)

There are many examples that fall into this category, especially for keywords that are less ambiguous and found in very specialized topics. For example "Skyrmion" has approxi-

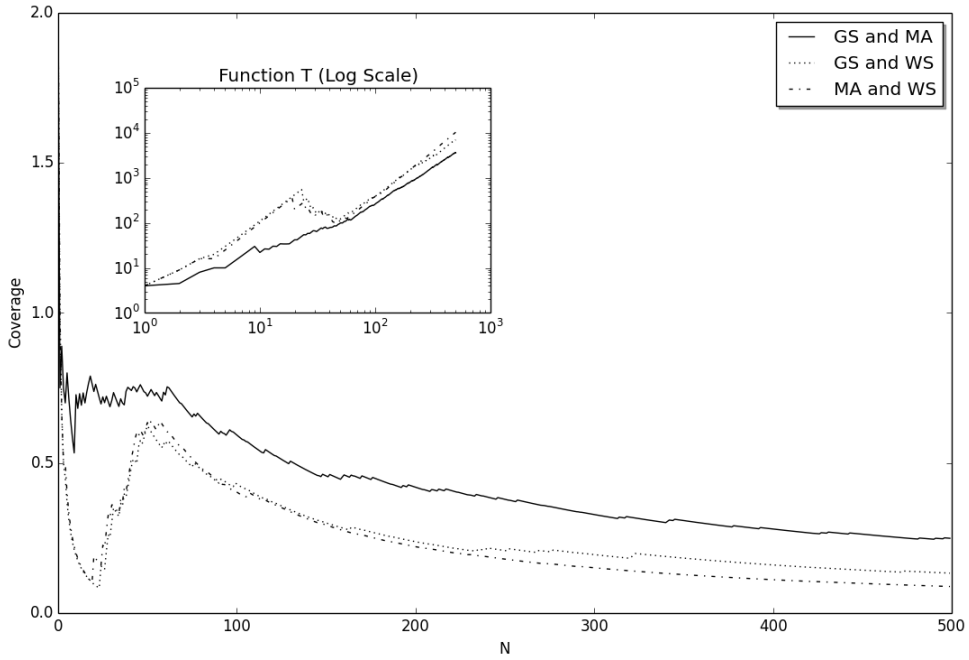

Figure 6.2: **Keyword: Kauffman Model**. At $n \approx 70$, the rate of change of coverage shifted from zero to negative. This implies one should stop around this point as further search has negative returns. An alternative stopping point is at $n \approx 20$ (local minimum) where it implies that the subsequent articles are already found by the other search engine.

mated 9000 articles in Google Scholar and most of the publications are also in the database of the other search engines. However every search engines have their own unique algorithms to rank the most relevant articles.

Fig. 6.3 shows that the results by the Web of Science initially deviates from Google Scholar and Microsoft Academic Search until $n \approx 100$ and $n \approx 180$ respectively. After which T converges for all pairwise comparisons. This implies that the initial ordering of "relevance" by Web of Science is partially the reverse of the other search engines.

More precisely after the local minimum, the subsequent articles by Web of Science are found in the earlier results of Google Scholar and Microsoft Academic Search. Therefore the coverage increases and there is little information gained.

Type IV (No Significant Feature)

There are many instances where the graphs do not fit into any of the above models. There

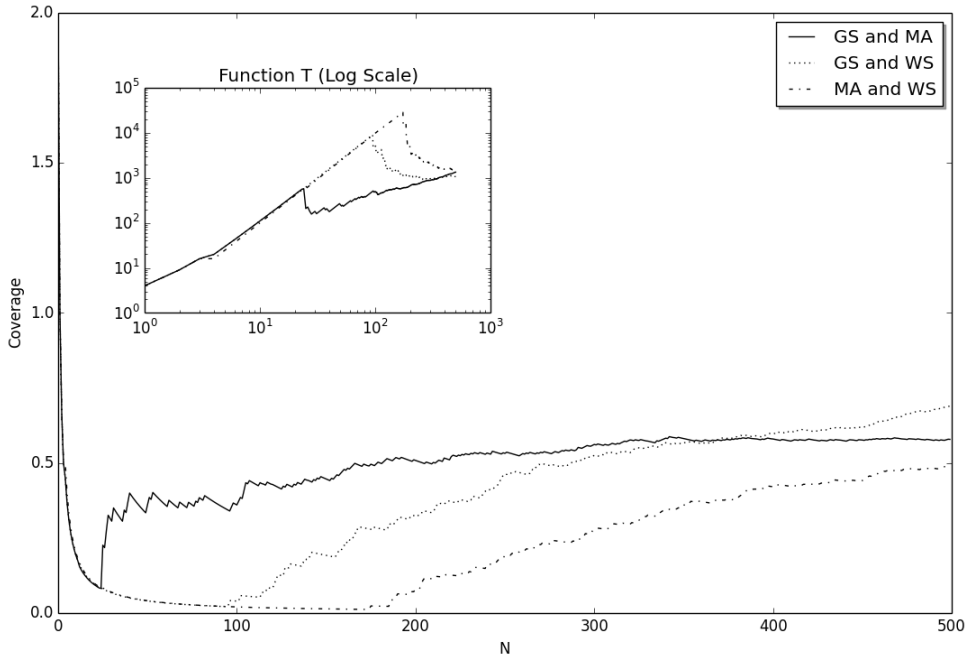

Figure 6.3: **Keyword: Skyrmion.** For Web of Science and Google Scholar, their local minimum is at $n \approx 100$. The subsequent articles in Web of Science matches the earlier articles in Google Scholar, vice versa.

is no significant minimum or maximum point for one to suggest a meaningful stop to the search. For example the solid line (Google Scholar versus Microsoft Academic Search) in Fig. 6.4 is the graph for "Causality Measures". There is no general rule to identify keywords that fall into this category.

6.4.2 Measure of Truncated-Ranking Similarities

The ordering from a search engine is determined by the relevance of the articles. For instance Google's algorithm has roots from Eigenvector Centrality where it ranks the quality of an article via the behavior of "word-of-mouth" recommendations. I.e. high ranking articles are either referred by other high ranking articles or by many independent articles.

Thus the growth of R in essence is also a measure of similarity for the centrality ranking of vertices. Specifically a linear R with slope 1 indicates high similarity while slow growing

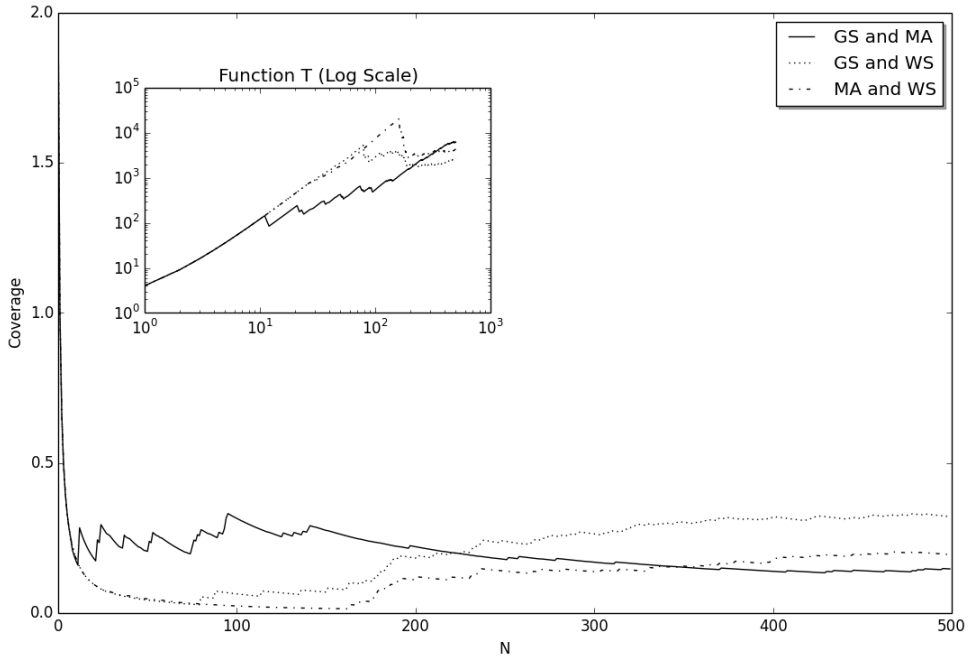

Figure 6.4: **Keyword: Causality Measures.** There is no significant reference point such that one can suggest a reasonable stop to the search.

(e.g. sublinear) indicates a lower degree of similarity. This is closely related to Spearman's Correlation and Kendall-tau Distance as ways to measure the similarity of ranked variables.

Spearman's Correlation is the variant of Pearson's Correlation for ranked variables where it measures how monotonically two rankings are related. Although it is relevant to our application, the model cannot be used for truncated dataset, i.e. comparing the top few elements of two rankings. Thus it is also not applicable for dynamical systems where the size of the network fluctuates and only the top centrality vertices are interesting.

Definition 6.4.1 ***(Kendall-tau Distance)****: Given two ordered sets* $X = \{x_1, \ldots, x_n\}$ *and* $Y = \{y_1, \ldots, y_n\}$*, the set of* n *observation is* $(x_1, y_1), \ldots, (x_n, y_n)$*. A pair of observations* (x_i, y_i) *and* (x_j, y_j) *are in agreement if both* $x_i > x_j$ *and* $y_i > y_j$ *or if both* $x_i < x_j$ *and* $y_i < y_j$*. The pair is in disagreement if* $x_i > x_j$ *and* $y_i < y_j$ *or if both* $x_i < x_j$ *and* $y_i > y_j$*.*

Hence the Kendall-tau Distance is:

$$\tau = \frac{(\textit{no. of agreement pairs}) - (\textit{no. of disagreement pairs})}{n(n-1)/2}. \tag{6.7}$$

Kendall-tau Distance measures how likely the order of two rankings agree. It handles truncated-ranking by ignoring elements that do not exists in both rankings. It is sensitive to the orderings of the elements and two rankings are independent (dissimilar) if they are random permutation of each other.

It is a good metric until one considers the size of the entire system. It is highly unlikely that in a large system that the top elements of two rankings are in common. Thus even though the ordering of two truncated-rankings might not agreeable in general, its effect is small relative to the fact that the elements are in common.

Squared Error as a Metric

The intuition of this metric is based on the observation that when two truncated-rankings are identical, R is a straight line with slope 1 intersecting zero. However when two truncated-rankings are totally dissimilar (no common elements), R is a straight line with slope 0.

Thus the similarity between two truncated-rankings is the Squared Error difference between R and the line with slope 1 intersecting zero. The smaller the Squared Error, the more similar two rankings are. This is based on the best fit line algorithm where the Squared Error between the data and line is minimized.

The maximum Squared Error is the difference between the lines $y = x$ and $y = 0$, hence to normalize the measure:

$$S = 1 - \frac{E(I,R)}{E(I,Z)}, \tag{6.8}$$

where for the top n elements, $I = \{1, 2, \ldots, n\}$ is the ideal case (y=x) and $Z = \{0, \ldots, 0\}$ (n zeros) is the case where there is no similarity. The Squared Error E is defined as:

$$E(X,Y) = \sum_{j=0}^{n} |x_j - y_j|^2. \tag{6.9}$$

Experiments Methodology

To simulate a dynamic network that varies in size, we construct a process that adds and removes random vertices from a network in each time step. Between each iteration, the Eigenvector Centrality of the vertices are computed and only the top 1000 vertices are compared. For example let G_t and G_{t+1} be the networks at time t and $t+1$ respectively. If Q_t and Q_{t+1} are the ordered lists of the top centrality vertices of G_t and G_{t+1} respectively, then R is derived by comparing Q_t and Q_{t+1} in the same way as we did with search engines.

Begin with a network on 10000 vertices constructed using Barabási-Albert's construction. In each iteration, x_r random vertices are removed and x_a vertices are added to the network where x_r and x_a are drawn from a normal distribution with mean 1000 and standard deviation 100. The new x_a vertices are added into the network using the same mechanism from Barabási-Albert's construction.

Empirical Results

Synthetic Network

Let Q_1 and Q_2 be two truncated-rankings on the index of vertices of a network. The similarity S has a mean of 0.8831 with standard deviation of 0.0697. It is highly correlated to the size of the set $Q_1 \cap Q_2$ with a Pearson's Coefficient of 0.984.

In contrast S is less correlated (Pearson's Coefficient of 0.2443) to Kendall-tau Distance as there are significant changes to the ordering of the top centrality vertices. More importantly the mean Kendall-tau Distance is 0.0332. This implies that Kendall-tau Distance claims that the two truncated-rankings are dissimilar. The reason is that there are many vertex pairs in one ranking that are not in the ranking of the other.

For example let $v_i, v_j \in Q_1$ where v_i is ranked higher than v_j in Q_1. Suppose $v_i \in Q_2$ and $v_j \notin Q_2$, then there is neither agreement nor disagreement between Q_1 and Q_2 on the pair (v_i, v_j). If there are many instances of such pairs, then the Kendall-tau Distance will be close to zero and implies that Q_1 and Q_2 are independent. However considering the size of the system, it would be unlikely to find many common top centrality vertices (e.g. v_i). Thus it is counter-intuitive to suggest that the two rankings are not similar.

Special Cases

Although $|Q_1 \cap Q_2|$ is highly correlated to our similarity metric S, this section presents some special cases of Q_1 and Q_2 to further distinguish S from the existing metrics.

Reverse Ranking: When Q_1 is the reverse of Q_2, $|Q_1 \cap Q_2| = 1$ and $S = 0.7492$. It will be particularly strange to state that the two truncated-rankings are identical given that $|Q_1 \cap Q_2| = 1$. Therefore our similarity metric distinguishes itself from the naive approximation of $|Q_1 \cap Q_2|$ by considering the order of the elements in the rankings.

Random Permutation: Let Q_1 be a random permutation of Q_2 and as before it is strange that both truncated-rankings are identical since $|Q_1 \cap Q_2| = 1$. Whereas from 1000 trials, the mean value of S and Kendall-tau Distance is 0.8993 and -0.0016 respectively. Thus S is able to encapsulate the significance of the orderings in the rankings.

Asymmetry: Our metric places more emphasis on the top positions of the truncated-ranking. For example let $Q_1 = \{v_a, v_b, \ldots, v_y, v_z\}$, $Q_2 = \{v_b, v_a, \ldots, v_y, v_z\}$ and $Q_3 = \{v_a, v_b, \ldots, v_z, v_y\}$ where the "..." is identical for all three truncated-rankings. For the other measures, the similarity between (Q_1, Q_2) is the same as the similarity of (Q_1, Q_3). However S shows that (Q_1, Q_2) is less similar than (Q_1, Q_3).

Let $|Q_1 \cap Q_2| = |Q_1|/2 = |Q_2|/2$ where the first halves of Q_1 and Q_2 are random permutations of each other. Thus there is no common element between the second halves of Q_1 and Q_2. From 1000 trials, the mean score for S and Kendall-tau Distance is 0.8629 and 0.7523 respectively.

If the situation is reversed, i.e. there is no common element between the first halves of Q_1 and Q_2, and the second halves are random permutation of each other, then the mean score of S and Kendall-tau Distance is 0.3616 and 0.7519 respectively. Since Kendall-tau Distance just counts the number of agreement/disagreement to the element pairs, it does not matter if the missing elements are positioned at the beginning or the end of the ranking. This is different from S as the agreement of the top rankings is more important the agreement at the bottom of the rankings.

Real World Data

The observation from our real world data (results from search engine) is similar to the results with the synthetic network in the previous experiments. Specifically our metric is

positively correlated to the size of $|Q_1 \cap Q_2|$ with a Pearson's Coefficient of > 0.95 for all pairwise comparisons of the search engines. In addition our metric is almost independent to Kendall-tau Distance with Pearson's Coefficient ≈ -0.1.

However it is the absolute score of the metrics that is particularly interesting. For instance between the search results of Google Scholar and Microsoft Academic Search, their mean similarity score for $|Q_1 \cap Q_2|$ and Kendall-tau Distance are 0.2799 and 0.0068 respectively. This implies that their results are not similar by those measures. In contrast, S has a score of 0.464 with standard deviation of 0.2347.

Since the score is normalized between 0 and 1, suppose we let the arbitrary threshold between similarity and dissimilarity to be 0.5. Thus our metric suggests that there is a huge variation between the closeness of the results of Google Scholar and Microsoft Academic Search. This supports the diverging conclusions from other empirical studies that they are *both* similar and dissimilar in general. Therefore our metric is normalized in a way such that it is good for measuring truncated-rankings like search engines' results.

6.5 Summary

Mark-And-Recapture is a methodology in population biology that can be used to estimate the breadth of a research from its literature reviews. However since prominent papers are more likely to be cited by others, the estimate is only sufficient to bound the minimum number of relevant literature. This gives one a general overview of the topic and to evaluate the completeness of his research.

This is particularly useful to assess disparate research like generalized graph models where there are many concurrent research from different disciplines. Furthermore it is also coincident that two independent literature reviews by Kivelä *et al.* and Boccaletti *et al.* were published around the same time.

This chapter summarizes one of the main research challenges of this thesis and the solutions derived to overcome them. It describes the process to improve the completeness of this thesis's bibliography and to quantify the time spent to consolidate the relevant research materials. Thus there is relevance to compile these ideas as a general research methodology despite that this chapter is slightly off tangent to the title of this thesis.

Chapter 7

End Notes

7.1 Summary

Interval graphs and multiplexes are network fine structures to encapsulate the relational properties of a system. Interval graph is well studied in Graph Theory, but the intractability to compute the boxicity limits its potential in many applications. On the other hand, multiplex is prevalent in many fields of science, but its literature is disparate and unorganized.

Chapter 3 presents the structural properties that was observed during the exploratory phase. Not only it describes their topologies, it helps us to draw the boundaries and map the direction for future work. Unlike many studies that look forward to formalize ways to understand and measure multiplexes [13, 30, 31, 33, 107], the chapter looks back at networks to hypothesize the types of relationships it *might* have. For example does the projection of a multiplex with Barabási-Albert and Watts-Strogatz graphs statistically similar to a scale-free network that exhibits high clustering coefficient?

The problem with current approaches to formalize multiplex metrics is that there is a huge degree of freedoms for subjective inputs from anecdotal perspectives. For instance in the ideal case, the different definitions of a multiplex-community are pretty similar conceptually. However as we deviate from the ideal situations (as in real-world systems), the solutions deviate and the opinions can be inconsistent (chapter 4).

Therefore in chapter 5 we look into interval graphs as a way formalize multiplex more objectively. For example the boxicity of the multiplex's projection determines the minimum

number of metrics to measure the vertices such that it describes the system. For simplicity we always want to minimize the number of relationships while preserving information, hence by occam's razor principle there are little reason to opt for a multiplex with more relationships than the hyperbox representation of its projection.

Finally to illustrate the challenges in compiling the disparate literature, chapter 6 describes a scientific process to organize the material. The novelty was the application of Mark-and-Recapture from population biology to gauge the completeness of this study and extends the idea as a quantitative measure for research.

7.2 Main Contributions in a Nutshell

This section is similar to the introduction (section 1.4) except that it is paraphrased differently as the technical details are now better understood. Hence this section will highlight this thesis's contributions to a greater context in Network Science and other applications.

The contributions of this research are disparate and easily lost within the text of this thesis. The reason is that there are many challenges and gaps in our current understanding of multiplexes that we have to bring in many somewhat unrelated ideas from other disciplines. Therefore in reading this thesis, the contributions are interwoven between prior research so that the flow of the writing is smooth and coherent. Hence the following lists down the contributions in a clear and concise manner:

- Chapter 3 is mainly some analytical properties of the resultant graph when we project multiplexes on two relationships. All the materials are original except for Theorem 3.5.1, 3.5.2 and 3.6.1.
- There are multiple contributions in Chapter 4:
  - The literature review on multiplex communities detection is more comprehensive and supplements the reviews by Boccaletti *et al.* and Kivelä *et al.*.
  - Derived new analytical bounds (corollary 4.3.2 and 4.3.3) for multiplex communities detection algorithms using probabilistic methods.
  - Structured Synthetic Random Multiplex is a new benchmark multiplex to highlight the multiple different definitions of communities in the same multiplex.

  - Showed that all the proposed multiplex communities detection algorithms are similar conceptually in ideal situations, but empirically very different as there are many non-ideal cases of a multiplex-community (Section 4.7).

- Another challenge in creating benchmarks for multiplex communities detection algorithms is to determine the number of relationships in a multiplex. Many current literature worked around this issue by prudent decisions and qualitatively argue their choices. However this research proposed an objective guideline using boxicity, a graph theory research that was active around 50 years ago (Section 5.1).

- Chapter 5 are some of the new work done on boxicity. For example in Section 5.1 we applied a heuristic method from Network Science (i.e. community detection to modularize the problem) to improve the computation hardness of boxicity. More importantly this heuristic method is more tolerant to experimental errors.

- Chapter 6 is a quantitative way to support the completeness of this thesis's bibliography using Mark-and-Recapture methods from population biology. For example we are able to show the materials that are lacking in the reviews by Boccaletti *et al.* and Kivelä *et al.* (Section 6.3.2). Since this is a generic method, its application can be easily extended to other research topics.

- Finally the additional novelty in Chapter 6 is to apply Mark-and-Recapture on search engines to determine a stopping rule for research, i.e. how many entries in the search results must we explore before there is diminishing return in knowledge gained.

7.3 Perspectives

Introspectively it is critical to question the necessity of these fine structures in practice, i.e. is it insightful to trade off simplicity for the additional relational information. It appears that in some situations the pursuit of high dimensionality in networks is like forcing a square peg in a round hole. The fact that it is mathematically coherent does not imply that it is conceptually significant. As a saying goes: "for every debate in science there is an isomorphic debate in the methodology of science" [82].

There is a degree of subjectivity when we translate a real-world problem into a multi-relational mathematical model. For example unlike a transportation multiplex where the relationships are well defined by the physical infrastructure, the relationships in a social systems are often *assumed* by the researchers' subjective understandings. Hence the correctness of the model is hard to verify. This is diametrically opposed to the objectivity and the issue is beyond the scientific inquiry [125].

This implies that the validity of a multi-relational model can be just as subjective and one is not erroneous to claim that all the models are useful in their own way. I.e. "*all models are wrong; the practical question is how wrong do they have to be to not be useful.*" [27]. Although it demonstrates the rich perspectives of the model, it can also lead one to reach conclusions to support his self-confirmation bias.

More importantly there are very little we can learn from the previous studies. For instance the formalism (e.g. multiplex-communities) of a social multiplex do not necessarily share the same ideas as a transportation multiplex. Hence the literature tends to be a collection of ad hoc methods that do not have a unifying theme or connection.

This is not a vitriol on the current literature, but to highlight how different its workings are from many scientific fields. The nature of multiplex research does not "stand on the shoulders of giants" such that the problems can be built upon the previous ideas. This is analogous to how combinatorics is placed along with the rest of mathematics [74].

The analogy is that multiplex research in general is a problem-solving discipline rather than theory-building. This is similar to combinatorics where typically the research do not to have a long chain of logical dependences [74]. This awareness helps to direct multiplex research in a more meaningful manner. I.e. one should emphasize on solving individual applications as a self-contained problem, rather than establishing a grand unified theory or a general-purpose tool for all multiplexes.

Therefore the conclusion is that multiplex research is still at its infancy, where there are many gaps in our understanding to apply it readily in practice. Specifically every application requires rigorous evaluation from its formalism to the tools for the analysis. Ergo the potential errors from the additional assumptions in multiplexity easily outweigh its benefits. Therefore there is a need to tie up some of the loose ends of the research with mathematical models like interval graphs that depends on different or less assumptions.

Glossary

Adjacent Vertices u and v are adjacent if they are in the edge set E, i.e. $(u, v) \in E$. 29

Asteroid-Triple Three independent vertices where every pair are connected by a path avoiding the neighbourhood of the third. 18

Barabási-Albert A scale-free graph generated by the preferential attachment mechanism. 33

Boxicity The minimum dimension to embed a graph as an intersection of axis-parallel boxes. 20

Centrality A metric of the most important vertices. 53

Chordal A graph with no cycle of length greater than 3. 18

Chromatic number The minimum number of colors needed to color the vertices such that no adjacent vertices have the same color. 27

Clique A complete subgraph. 19

Clustering coefficient Average ratio of all vertex pairs who are neighbors of each other to all pairs of neighbors. 46

Community A dense cluster of vertices that is sparsely connected to the rest of the graph. 30

Complement graph The complement of a graph $G(V, E)$ is the graph where its edges is the set of all possible edges that are not in E. 72

Complete A graph is complete if all vertex pairs are adjacent. 58

Component A connected subgraph. 78

Connected There exists a path between all vertex pairs. 32

Cycle A closed path of distinct edges with the same starting and ending vertex. 30

Degree The number of edges incident to the vertex. 39

Edge An unordered pair of vertices. 12

Erdős-Rényi A random graph where vertex pairs are connected with fixed probability. 32

Graph A pair of disjoint sets $G(V, E)$ where $E \subseteq V \times V$. V and E are known as the vertex set and the edge set respectively. 12

Independent A set of non adjacent vertices. 18

Interval Graph A duality of a graph and a set of intervals, where overlapping intervals denote the adjacency of the vertices. 18

Loop An edge that connects a vertex to itself. 132

Louvain Algorithm An optimization algorithm to partition a graph and maximizes its modularity. 34

Modularity A metric to measure how different the communities are from a random graph. 34

Multiplex A set of graphs on the same vertex set. 28

Neighbors The the set of all adjacent vertices. 18

Network see graph. 12

Overlap Edges $e(u, v)$ and $e'(u', v')$ overlap if $u = u'$ and $v = v'$. 36

Parallel edge A pair of vertices with multiple edges between them. 132

Path A sequence of edges connecting a set of vertices. 18

Preferential attachment The process where new vertices are more likely to be adjacent to higher degree vertices. 33

Projection The graph from the union of the edge sets in a multiplex. 29

Regular A graph where all the vertices are of the same degree. 35

Ring Lattice A cycle of vertices. 33

Scale-free The property that a graph's degree distribution follows a power-law. 130

Simple graph An undirected graph with no loop or parallel edge. 33

Small-world The property that a high clustering coefficient graph has average path length that grows proportionally to the logarithm of the size of the graph. 132

Star graph A tree on n vertices with one vertex having degree $n-1$ and the rest of the vertices having degree 1. 37

Subgraph A subgraph of graph G is a graph whose vertex set or edge set is a subset of those in G. 130

Tree A connected simple graph with no cycles. 132

Vertex Fundamental unit of a graph. 12

Watts-Strogatz A small-world graph constructed by randomly rewiring the edges of a lattice. 33

References

[1] A. Adiga, L.S. Chandran, and N. Sivadasan. Lower bounds for boxicity. *Combinatorica*, 2014.

[2] N. Alon and J.H. Spencer. *The probabilistic method*. John Wiley & Sons, 2004.

[3] D. Ambrosino and A. Sciomachen. Hub locations in urban multimodal networks. *European Transport Trasporti Europei*, (51), 2012.

[4] K. Andreev and H. Racke. Balanced graph partitioning. *Theo. of Comp. Sys.*, 2006.

[5] S. A. Applin and M. Fischer. Everybody is talking to each other without talking to each other. *Annual Meeting of the Amer. Anthropological Ass.*, 2012.

[6] Tomaso Aste, Ruggero Gramatica, and T Di Matteo. Exploring complex networks via topological embedding on surfaces. *Physical Review E*, 86(3):036109, 2012.

[7] N. TJ Bailey. *The mathematical theory of infectious diseases and its applications*. Charles Griffin & Company Ltd., 1975.

[8] C. Balbuena and M.I. Ortego. The distribution of extremes in the degree sequence: A gumbel distribution approach. *Appl. Math. Lett.*, 22(4):553–556, 2009.

[9] K.T. Balińska. The random f-graph process. *Annals of Discrete Math.*, 55, 1993.

[10] A.L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[11] A. Barrat and M. Weigt. On the properties of small-world network models. *The Euro. Phys. J.l B*, 13(3):547–560, 2000.

[12] F. Barzinpour, B.H. Ali-Ahmadi, S. Alizadeh, S. Golamreza, and J. Naini. Clustering networks heterogeneous data in defining a comprehensive closeness centrality index. *Mathematical Problems in Engineering*, 2014(202350), 2014.

[13] F. Battiston, V. Nicosia, and V. Latora. Structural measures for multiplex networks. *Phys. Rev. E*, 89:032804, Mar 2014.

[14] D.A. Bennett, N.K. Latham, C. Stretton, and C.S. Anderson. Capture-recapture is a potentially useful method for assessing publication bias. *J. of clinical epidemiol.*, 2004.

[15] S. Benzer. On the topology of the genetic fine structure. *Proc. of the Nat. Acad. of Sci.*, 45(11):1607, 1959.

[16] M. Berlingerio, M. Coscia, and F. Giannotti. Finding and characterizing communities in multidimensional networks. In *ASONAM*, pages 490–494. IEEE, 2011.

[17] M. Berlingerio, M. Coscia, F. Giannotti, A. Monreale, and D. Pedreschi. Multidimensional networks: foundations of structural analysis. *World Wide Web*, 2012.

[18] M. Berlingerio, F. Pinelli, and F. Calabrese. Abacus: frequent pattern mining-based community discovery in multidimensional networks. *DMKD*, 27(3), 2013.

[19] G. Bianconi. Statistical mechanics of multiplex networks: Entropy and overlap. *Phys. Rev. E*, 87:062806, 2013.

[20] T. Biedl, B. Brejová, E.D. Demaine, A.M. Hamel, A. López-Ortiz, and T. Vinař. Finding hidden independent sets in interval graphs. *Theo. Comp. Sci.*, 2004.

[21] S. Bikhchandani, D. Hirshleifer, and I. Welch. A theory of fads, fashion, custom, and cultural change as informational cascades. *J. of political Economy*, 1992.

[22] V.D. Blondel, J.L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *J. of Statistical Mechanics*, (10), 2008.

[23] S. Boccaletti, G. Bianconi, R. Criado, C.I. del Genio, J. Gómez-Gardeñes, M. Romance, I. Sendiña-Nadal, Z. Wang, and M. Zanin. The structure and dynamics of multilayer networks. *Phys. Rep.*, 2014.

[24] B. Bollobás. *Random graphs*. Springer, 1998.

[25] A. Booth. How much searching is enough? comprehensive versus optimal retrieval for technology assessments. *Inter. J. of tech. assessment in health care*, 2010.

[26] K.S. Booth and G.S. Lueker. Testing for the consecutive ones property, interval graphs, and graph planarity using pq-tree algorithms. *J. of Comp. & Sys. Sci.*, 1976.

[27] G.E.P. Box and N.R. Draper. *Empirical Model Building and Response Surfaces*. John Wiley & Sons, 1987.

[28] P. Bródka, T. Filipowski, and P. Kazienko. An introduction to community detection in multi-layered social network. In *Info. Sys., E-learning, and Knowledge Management Research*, pages 185–190. Springer, 2013.

[29] P. Bródka and T. Grecki. mlfr benchamark: Testing community detection algorithms in multilayered, multiplex and multiple social networks. 2014.

[30] P. Bródka, P. Kazienko, K. Musiał, and K. Skibicki. Analysis of neighbourhoods in multi-layered dynamic social networks. *Inter. J. of Comp. Intelligence Sys.*, 2012.

[31] P. Bródka, K. Skibicki, P. Kazienko, and K. Musial. A degree centrality in multi-layered social network. In *Comp. Aspects of Social Net.* IEEE, 2011.

[32] G.J. Browne, M.G. Pitts, and J.C. Wetherbe. Cognitive stopping rules for terminating information search in online tasks. *MIS quarterly*, pages 89–104, 2007.

[33] F. Bry, F. Kneil, K.A. Weiand, and T. Furche. Term-specific eigenvector-centrality in multi-relation networks. *IJSNM*, 1(2):141–159, 2012.

[34] B.M. Bui-Xuan and N.S. Jones. How modular structure can simplify tasks on networks. *preprint arXiv:1305.4760*, 2013.

[35] B.M. Bui-Xuan and N.S. Jones. How modular structure can simplify tasks on networks: parameterizing graph optimization by fast local community detection. *Proc. of the Royal Society of London A*, 470(2170), 2014.

[36] D. Cai, Z. Shao, X. He, X. Yan, and J. Han. Community mining from multi-relational networks. In *Practice of Knowledge Discovery in Databases*, 2005.

[37] A. Cardillo, J. Gómez-Gardeñes, M. Zanin, M. Romance, D. Papo, F. del Pozo, and S. Boccaletti. Emergence of network features from multiplexity. *Sci. Rep.*, 2013.

[38] A. Cardillo, M. Zanin, J. Gómez-Gardeñes, M. Romance, A.J.G del Amo, and S. Boccaletti. Modeling the multi-layer nature of the european air transport network. *preprint arXiv:1211.6839*, 2012.

[39] M.C. Carlisle and E.L. Lloyd. On the k-coloring of intervals. *Discrete Appl. Math.*, 59(3):225–235, 1995.

[40] M.F. Cattin, L.F. Bersier, C. Banašek-Richter, R. Baltensperger, and J.P. Gabriel. Phylogenetic constraints and adaptation explain food-web structure. *Nature*, 427(6977):835–839, 2004.

[41] L.S. Chandran, A. Das, and C.D. Shah. Cubicity, boxicity, and vertex cover. *Discrete Math.*, 309(8):2488–2496, 2009.

[42] L.S. Chandran, M.C. Francis, and N. Sivadasan. Geometric representation of graphs in low dimension using axis parallel boxes. *Algorithmica*, 56(2):129–140, 2010.

[43] L.S. Chandran, R. Mathew, and N. Sivadasan. Boxicity of line graphs. *Discrete Mathematics*, 311(21):2359–2367, 2011.

[44] S.L. Chandran, Mathew C Francis, and Naveen Sivadasan. Boxicity and maximum degree. *J. of Combinatorial Theory, Series B*, 98(2):443–445, 2008.

[45] A.L. Chen, F.J. Zhang, and H. Li. The degree distribution of the random multigraphs. *Acta Mathematica Sinica, English Series*, 28(5):941–956, 2012.

[46] R. M. Christley, G. L. Pinchbeck, R. G. Bowers, D. Clancy, N. P. French, R. Bennett, and J. Turner. Infection in social networks: Using network analysis to identify high-risk individuals. *Amer. J. of Epidemiol.*, 162(10):1024–1031, 2005.

[47] J.E. Cohen. *Food webs and niche space*. Princeton Univ. Press, 1978.

[48] L. M. Collins and C. W. Dent. Omega: A general formulation of the rand index of cluster recovery suitable for non-disjoint solutions. *Multivariate Behavioral Research*, 23(2):231–242, 1988.

[49] G. Cormode and S. Muthukrishnan. Space efficient mining of multigraph streams. In *Principles of database systems*, pages 271–282. ACM, 2005.

[50] M.B. Cozzens. *Higher and Multi-dimensional Analogues of Interval Graphs*. Rutgers University, 1981.

[51] M.B. Cozzens and F.S. Roberts. Computing the boxicity of a graph by covering its complement by cointerval graphs. *Discrete Appl. Math.*, 6(3):217–228, 1983.

[52] E. Cozzo, A. Arenas, and Y. Moreno. Stability of boolean multilevel networks. *Phys. Rev. E*, 86:036115, 2012.

[53] E. Cozzo, M. Kivelä, M. De Domenico, A. Solé, A. Arenas, S. Gómez, M.A. Porter, and Y. Moreno. Clustering Coefficients in Multiplex Networks. *preprint arXiv:1307.6780*, 2013.

[54] M. De Domenico, A. Lancichinetti, A. Arenas, and M. Rosvall. Identifying modular flows on multilayer networks reveals highly overlapping organization in social systems. *arXiv:1408.2925*, 2014.

[55] M. De Domenico, V. Nicosia, A. Arenas, and V. Latora. Layer aggregation and reducibility of multilayer interconnected networks. *preprint arXiv:1405.0425*, 2014.

[56] Y.X. Dong, J. Tang, S. Wu, J.L. Tian, N.V. Chawla, J.H. Rao, and H.H. Cao. Link prediction and recommendation across heterogeneous social networks. *ICDM*, 2012.

[57] S.N. Dorogovtsev, J.F.F. Mendes, and A.N. Samukhin. Structure of growing networks with preferential linking. *Phys. Rev. Lett.*, 85(21):4633, 2000.

[58] D.M. Dunlavy, T.G. Kolda, and W.P. Kegelmeyer. Multilinear algebra for analyzing data with multiple linkages. *Graph Algo. in the Language of Linear Algebra*, 2011.

[59] N. Eagle and A.S. Pentland. Reality mining: sensing complex social systems. *Journal Personal and Ubiquitous Computing*, 10, 2006.

[60] A. Eklöf, U. Jacob, J. Kopp, J. Bosch, R. Castro-Urgal, N.P. Chacoff, B. Dalsgaard, C. Sassi, M. Galetti, P.R. Guimarães, et al. The dimensionality of ecological networks. *Ecology Lett.*, 16(5):577–583, 2013.

[61] P. Erdős and A. Rényi. *Publicationes Mathematicae*, 1959.

[62] L. Esperet and G. Joret. Boxicity of graphs on surfaces. *Graphs and Combinatorics*, 29(3):417–427, 2013.

[63] P.C. Fishburn. Interval graphs and interval orders. *Discrete Math.*, 55(2), 1985.

[64] S. Fortunato. Community detection in graphs. *Phys. Rep.*, 486(3), 2010.

[65] S. Fortunato and C. Castellano. Community Structure in Graphs. *preprint arXiv: 0712.2716*, 2007.

[66] L.C. Freeman. Spheres, cubes and boxes: Graph dimensionality and network structure. *Social Networks*, 5(2):139 – 156, 1983.

[67] M.R. Garey and D.S. Johnson. *Computers and Intractability; A Guide to the Theory of NP-Completeness*. W. H. Freeman & Co., 1990.

[68] E.A. Gattass and G.L. Nemhauser. An application of vertex packing to data analysis in the evaluation of pavement deterioration. *Operations Research Lett.*, 1(1), 1981.

[69] E.N. Gilbert. Random graphs. *The Annals of Math. Statistics*, 30(4), 1959.

[70] M. Girvan and M.E.J. Newman. Community structure in social and biological networks. *Proc. of the National Academy of Sciences*, 99(12):7821–7826, 2002.

[71] M. Gjoka, C.T. Butts, M. Kurant, and A. Markopoulou. Multigraph sampling of online social networks. *J. Sel. Areas Com. on Meas. of Internet Topol.*, 2011.

[72] P.W. Goldberg, M.C. Golumbic, H. Kaplan, and R. Shamir. Four strikes against physical mapping of dna. *J. of Computational Biology*, 2(1):139–152, 1995.

[73] J. Goldenberg, B. Libai, and E. Muller. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing Lett.*, 12(3), 2001.

[74] W.T. Gowers. The two cultures of mathematics. *Mathematics: frontiers and perspectives*, 65, 2000.

[75] M. Granovetter. Threshold models of collective behavior. *Amer. J. of sociol.*, 1978.

[76] R. Guns. *Missing links: Predicting interactions based on a multi-relational network structure with applications in informetrics*. Universiteit Antwerpen, 2012.

[77] J.J. Hao, S.M. Cai, Q.B. He, and Z.R. Liu. The interaction between multiplex community networks. *Chaos: An Interdisciplinary J. of Nonlinear Sci.*, 21(1), 2011.

[78] R.A. Harshman. Foundations of the PARAFAC procedure. *UCLA Working Papers in Phonetics*, 16(1):84, 1970.

[79] A.W. Harzing. Publish or perish. *http://www.harzing.com/pop.htm*, 2007.

[80] H.W. Hethcote. The mathematics of infectious diseases. *SIAM review*, 42(4), 2000.

[81] C. A. R. Hoare. Algorithm 64: Quicksort. *Commun. ACM*, 4(7):321–, July 1961.

[82] D.R. Hofstadter. *Metamagical Themas: Questing for the Essence of Mind and Pattern*. Basic Books, Inc., 1985.

[83] T. Hossmann, G. Nomikos, T. Spyropoulos, and F. Legendre. Collection and analysis of multi-dimensional network data for opportunistic networking research. *Computer Communications*, 35(13):1613 – 1625, 2012.

[84] L. Hubert and P. Arabie. Comparing partitions. *J. of classification*, 2(1), 1985.

[85] J.C. Jiang and M. Jaeger. Community detection for multiplex social networks based on relational bayesian networks. In *Foundations of Intelligent Systems*. 2014.

[86] F. Jordán and I. Scheuring. Network ecology: topological constraints on ecosystem dynamics. *Physics of Life Reviews*, 1(3):139–172, 2004.

[87] J.R. Jungck, G. Dick, and A.G. Dick. Computer-assisted sequencing, interval graphs, and molecular evolution. *Biosystems*, 15(3):259–273, 1982.

[88] H. Kaplan and R. Shamir. Bounded degree interval sandwich problems. *Algorithmica*, 24(2):96–104, 1999.

[89] H. Kaplan, R. Shamir, and R.E. Tarjan. Tractability of parameterized completion problems on chordal, strongly chordal, and proper interval graphs. *SICOMP*, 1999.

[90] M. Kastner, S. Straus, and C.H. Goldsmith. Estimating the horizon of articles to decide when to stop searching in systematic reviews. In *AMIA*, page 389, 2007.

[91] M. Kastner, S.E. Straus, K. McKibbon, and C.H. Goldsmith. The capture–mark–recapture technique can be used as a stopping rule when searching in systematic reviews. *J. of clinical epidemiol.*, 62(2):149–157, 2009.

[92] P. Kazienko, K. Musial, E. Kukla, T. Kajdanowicz, and P. Bródka. Multidimensional social network: model and analysis. In *Comp. Collective Intelligence*. 2011.

[93] D. Kempe, J. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In *Inter. conf. on Knowledge discovery and data mining*, 2003.

[94] J.R. Kettenring. Canonical analysis of several sets of variables. *Biometrika*, 58(3):pp. 433–451, 1971.

[95] M. Kivelä, A. Arenas, M. Barthelemy, J.P. Gleeson, Y. Moreno, and M.A. Porter. Multilayer networks. *J. of Complex Networks*, 2014.

[96] A.W.J. Kolen, J.K. Lenstra, C.H. Papadimitriou, and F.C.R. Spieksma. Interval scheduling: A survey. *Naval Research Logistics*, 54(5):530–543, 2007.

[97] J. Komlós, A. Shokoufandeh, M. Simonovits, and E. Szemerédi. The regularity lemma and its applications in graph theory. In *Theo. Aspects of Comp. Sci.* 2002.

[98] V. Koponen. Random graphs with bounded maximum degree: asymptotic structure and a logical limit law. *Discrete Math. and Theo. Comp. Sci.*, 14(2), 2012.

[99] D. Kühn and D. Osthus. Maximizing several cuts simultaneously. *Combinatorics, Probability & Computing*, 16(2):277–283, 2007.

[100] Renaud Lambiotte. Multi-scale modularity in complex networks. In *Modeling and optimization in mobile, ad hoc and wireless networks (WiOpt), 2010 Proceedings of the 8th International Symposium on*, pages 546–553. IEEE, 2010.

[101] A. Lancichinetti, S. Fortunato, and J. Kertész. Detecting the overlapping and hierarchical community structure in complex networks. *New J. of Phys.*, 2009.

[102] D. Lane, J. Dykeman, M. Ferri, C.H. Goldsmith, and H.T. Stelfox. Capture-mark-recapture as a tool for estimating the number of articles available for systematic reviews in critical care medicine. *J. of critical care*, 28(4):469–475, 2013.

[103] L.J. LeBlanc. Transit system network design. *Transportation Research Part B: Methodological*, 22(5):383–390, 1988.

[104] K.M. Lee, J.Y. Kim, W.K. Cho, I.K. Goh, and I.M. Kim. Correlated multiplexity and connectivity of multiplex random networks. *New J. of Physics*, 14(3), 2012.

[105] M. Leginus, P. Dolog, and V. emaitis. Improving tensor based recommenders with clustering. In *User Modeling, Adaptation, and Personalization*. 2012.

[106] C. Lekkeikerker and J. Boland. Representation of a finite graph by a set of intervals on the real line. *Fundamenta Mathematicae*, 51(1):45–64, 1962.

[107] F. Lerch, J. Sydow, and K.G. Provan. Cliques within clustersmulti-dimensional network integration and innovation activities. In *Annual Colloquium of the European Group for Organizational Studies*, 2006.

[108] C. Li, J. Luo, J.Z.W. Huang, and J.P. Fan. Multi-layer network for influence propagation over microblog. In *Intelligence and Security Informatics*. 2012.

[109] R.N. Lichtenwalter. *Network analysis and link prediction: Effective and meaningful modeling and evaluation*. University of Notre Dame, 2012.

[110] C.W. Loe and H.J. Jensen. Bibliographic search with mark-and-recapture. *Physica A: Statistical Mechanics and its Applications*.

[111] C.W. Loe and H.J. Jensen. Comparison of communities detection algorithms for multiplex. *Physica A: Statistical Mechanics and its Applications*.

[112] C.W. Loe and H.J. Jensen. Revisiting interval graphs for network science. *arXiv:1503.07199*.

[113] C.W. Loe and H.J. Jensen. Centrality of unions of networks on the same vertex set. *arXiv:1309.6629*, 2013.

[114] C.W. Loe and H.J. Jensen. Edge union of networks on the same vertex set. *J. of Phys. A*, 46(24):245002, 2013.

[115] A. Louati, J.E. Haddad, and S. Pinson. Trust-based service discovery in multi-relation social networks. In *ICSOC*, volume 7636 of *LNCS*, 2012.

[116] W.F. Lu and W.L. Hsu. A clustering algorithm for interval graph test on noisy data. In *Experimental and Efficient Algorithms*, pages 195–208. Springer, 2003.

[117] M. Magnani and L. Rossi. Formation of multiple networks. In *Social Computing, Behavioral-Cultural Modeling and Prediction*, pages 257–264. Springer, 2013.

[118] C.D. Manning, P. Raghavan, and H. Schütze. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge, 2008.

[119] T.A. McKee and F. R. McMorris. *Topics in Intersection Graph Theory*. Monographs on Discrete Mathematics and Applications. Soc. for Ind. and Appl. Math., 1999.

[120] A. Mirzal and M. Furukawa. Node-Context Network Clustering using PARAFAC Tensor Decomposition, 2010.

[121] N. Miyoshi, T. Shigezumi, R. Uehara, and O. Watanabe. Scale free interval graphs. In *Algo. Aspects in Information and Management*, pages 292–303. 2008.

[122] D. Mouillot, B.R. Krasnov, and R. Poulin. High intervality explained by phylogenetic constraints in host-parasite webs. *Ecology*, 89(7):2043–2051, 2008.

[123] P.J. Mucha, T. Richardson, K. Macon, M.A. Porter, and J.P. Onnela. Community structure in time-dependent, multiscale, and multiplex networks. *Science*, 328(5980):876–878, 2010.

[124] S.A. Myers, C.G. Zhu, and J. Leskovec. Information diffusion and external influence in networks. In *Inter. conf. on Knowledge discovery and data mining*, 2012.

[125] T. Nagel. What is it like to be a bat? *The philosophical rev.*, 1974.

[126] A. Nagurney. A multiclass, multicriteria traffic network equilibrium model. *Math. and Comp. Modelling*, 32(34):393 – 411, 2000.

[127] A. Natanzon, R. Shamir, and R. Sharan. A polynomial approximation algorithm for the minimum fill-in problem. *SIAM J. on Comp.*, 30(4):1067–1079, 2000.

[128] N. Nefedov. Multiple-membership communities detection in mobile networks. In *WIMS*, page 64. ACM, 2011.

[129] M.E.J. Newman. Detecting community structure in networks. *The Euro. Phys. J. B*, 38(2):321–330, 2004.

[130] M.E.J. Newman. Modularity and community structure in networks. *Proc. of the Nat. Acad. of Sci.*, 103(23):8577–8582, 2006.

[131] V. Nicosia, G. Bianconi, V. Latora, and M. Barthelemy. Growing multiplex networks. *Phys. Rev. Lett.*, 111:058701, Jul 2013.

[132] V. Nicosia, G. Bianconi, V. Latora, and M. Barthelemy. Nonlinear growth and condensation in multiplex networks. *Phys. Rev. E*, 90:042807, Oct 2014.

[133] J.F. Padgett and C.K. Ansell. Robust action and the rise of the medici, 1400-1434. *Amer. J. of Sociol.*, 98(6):1259–1319, 1993.

[134] E.E. Papalexakis, L. Akoglu, and D. Ienco. Do more views of a graph help? community detection and clustering in multi-graphs. *Inter. Conf. on Info. Fusion*, 2013.

[135] V. Patel. Cutting two graphs simultaneously. *J. of Graph Theory*, 57(1):19–32, 2008.

[136] D.M. Pennock, G.W. Flake, S. Lawrence, E.J. Glover, and C.L. Giles. Winners don't take all. *Proc. of the Nat. Acad. of Sci*, 99(8):5207–5211, 2002.

[137] M.A. Porter, J.P. Onnela, and P.J. Mucha. Communities in networks. *Notices of the AMS*, 56(9):1082–1097, 2009.

[138] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi. Defining and identifying communities in networks. *Proc. of the Nat. Acad. of Sci.*, 101(9), 2004.

[139] J.M.O. Ranola, S. Ahn, M.E. Sehl, D.J. Smith, and K. Lange. A poisson model for random multigraphs. *Bioinformatics*, 26(16), 2010.

[140] F. S. Roberts. On the boxicity and cubicity of a graph. *Recent Progress in Combinatorics*, pages 301–310, 1996.

[141] N. Roberts and S.F. Everton. Roberts and everton terrorist data: Noordin top terrorist network (subset). *url:sites.google.com/site/sfeverton18/research/appendix-1*, 2011.

[142] K.A. Robinson, A.G. Dunn, G. Tsafnat, and P. Glasziou. Citation networks of related trials are often disconnected. *J. of clinical epidemiol.*, 2014.

[143] M.A. Rodriguez and J. Shinavier. Exposing multi-relational networks to single-relational network analysis algorithms. *J. of Informetrics*, 4(1):29 – 41, 2010.

[144] M.G. Rodriguez, J. Leskovec, and A. Krause. Inferring networks of diffusion and influence. In *Inter. conf. on Knowledge discovery and data mining*, 2010.

[145] V. Rosato, L. Issacharoff, F. Tiriticco, S. Meloni, S. Porcellinis, and R. Setola. Modelling interdependent infrastructures using interacting dynamical models. *Inter. J. of Critical Infrastructures*, 4(1):63–79, 2008.

[146] G. Rossetti, M. Berlingerio, and F. Giannotti. Scalable link prediction on multidimensional networks. In *ICDM Workshops*, pages 979–986, 2011.

[147] M. Rosvall, D. Axelsson, and C.T. Bergstrom. The map equation. *The Euro. Phys. J.*, 178(1):13–23, 2009.

[148] S.E. Schaeffer. Graph clustering. *Comp. Sci. Rev.*, 1(1):27–64, 2007.

[149] E.R. Scheinerman. *Intersection Classes and Multiple Intersection Parameters of Graphs*. Princeton University, 1984.

[150] E.R. Scheinerman. An evolution of interval graphs. *Discrete Math.*, 82(3), 1990.

[151] Z.E. Schnabel. The estimation of total fish population of a lake. *American Mathematical Monthly*, pages 348–352, 1938.

[152] A. Sela, H. Oved, and I. Ben-Gal. Information spread in a connected world. In *Proc. of Collective Intelligence 2014*. MIT Boston, 2014.

[153] J. Skvoretz and F. Agneessens. Reciprocity, multiplexity, and exchange: Measures. *Quality & Quantity*, 41(3):341–357, 2007.

[154] R. Solomonoff and A. Rapoport. Connectivity of random nets. *The bulletin of mathematical biophysics*, 13(2):107–117, 1951.

[155] S. Soundarajan and J. Hopcroft. Using community information to improve the precision of link prediction methods. In *Inter. conf. companion on WWW*, 2012.

[156] T.R.E. Southwood and P.A. Henderson. *Ecological Methods*. Wiley, 2009.

[157] H.T. Stelfox, G. Foster, D. Niven, A.W. Kirkpatrick, and C.H. Goldsmith. Capture-mark-recapture to estimate the number of missed articles for systematic reviews in surgery. *The American J. of Surgery*, 206(3):439–440, 2013.

[158] D.B. Stouffer, J. Camacho, and L.A.N. Amaral. A robust measure of food web intervality. *Proc. of the Nat. Acad. of Sci.*, 103(50):19015–19020, 2006.

[159] M. Szell, R. Lambiotte, and S. Thurner. Multirelational organization of large-scale social networks in an online world. *Proc. of the Nat. Acad. of Sci.*, 107(31), 2010.

[160] E. Szemerédi. Regular partitions of Graphs. In *Colloques Internationaux C.N.R.S 260 - Problème Combinatoire et Thèorie des Graphes, Orsay*, 1976.

[161] L. Tang, X.F. Wang, and H. Liu. Uncovering groups via heterogeneous interaction analysis. *Inter. Conf. on Data Mining*, 2009.

[162] C. Thomassen. Interval representations of planar graphs. *J. of Combinatorial Theory, Series B*, 40(1):9–20, 1986.

[163] T.W. Valente, K. Coronges, C. Lakon, and E. Costenbader. How correlated are network centrality measures? *Connections (Toronto, Ont.)*, 28(1), 2008.

[164] D.J. Watts and S.H. Strogatz. Collective dynamics of small-world networks. *nature*, 393(6684):440–442, 1998.

[165] Z.A. Wu, W.P. Yin, J. Cao, G.D. Xu, and A. Cuzzocrea. Community detection in multi-relational social networks. In *Web Info. Sys. Eng.* 2013.

[166] Z.M. Yin and S.S. Khaing. Multi-layered graph clustering in finding the community cores. *Inter. J. of Adv. Research in Comp. Eng. & Tech.*, 2, 2013.

[167] W. W. Zachary. An information flow model for conflict and fission in small groups. *J. of Anthropological Research*, 33:452–473, 1977.

[168] G.Y. Zhu and K. Li. A unified model for community detection of multiplex networks. In *Web Info. Sys. Eng.* 2014.