TWO PROBLEMS ON THE STRUCTURE-IDENTITY RELATIONSHIP ON NETWORKS

Jefferson Elbert Simões

Tese de Doutorado apresentada ao Programa de Pós-graduação em Engenharia de Sistemas e Computação, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Doutor em Engenharia de Sistemas e Computação.

Orientador: Daniel Ratton Figueiredo

Rio de Janeiro
Setembro de 2016

TWO PROBLEMS ON THE STRUCTURE-IDENTITY RELATIONSHIP ON NETWORKS

Jefferson Elbert Simões

TESE SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE DOUTOR EM CIÊNCIAS EM ENGENHARIA DE SISTEMAS E COMPUTAÇÃO.

Examinada por:

Prof. Daniel Ratton Figueiredo, Ph.D.

Prof. Valmir Carneiro Barbosa, Ph.D.

Prof. Franklin de Lima Marquezino, D.Sc.

Prof[a]. Maria Eulália Vares, Ph.D.

Prof. Daniel Sadoc Menasché, Ph.D.

Prof. Augusto Quadros Teixeira, Ph.D.

RIO DE JANEIRO, RJ – BRASIL
SETEMBRO DE 2016

Simões, Jefferson Elbert

Two problems on the structure-identity relationship on networks/Jefferson Elbert Simões. – Rio de Janeiro: UFRJ/COPPE, 2016.

XI, 113 p.: il.; 29, 7cm.

Orientador: Daniel Ratton Figueiredo

Tese (doutorado) – UFRJ/COPPE/Programa de Engenharia de Sistemas e Computação, 2016.

Referências Bibliográficas: p. 105 – 113.

1. Identidade estrutural. 2. Redes aleatórias. 3. Modelo de Erdős-Rényi. 4. Probabilidade. 5. Simetria em grafos. 6. Simetria local. 7. Casamento de grafos. 8. Casamento múltiplo de grafos. 9. Hiperpermutação. I. Figueiredo, Daniel Ratton. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia de Sistemas e Computação. III. Título.

"Sometimes,
when the spirit's left alone,
we must believe in something
to find if we've grown.
[...]
Sometimes,
a view from sinless eyes
centers our perspective
and pacifies our cries."
(John Petrucci)

Agradecimentos

Agradeço à minha mãe, Celia, ao meu pai, Celso, e ao meu irmão, Douglas, por serem meu porto seguro, por todo amor e apoio incondicional em todos os anos de minha formação. Eu não teria sido capaz de trilhar todo este caminho sem a confiança que vocês depositaram em mim e sem todas as lições de vida que aprendi e ainda aprendo com vocês. Agradeço também à minha noiva, Vivien, por ter transformado minha vida em tão pouco tempo que estamos juntos. Jamais imaginei aprender tanto com alguém tão semelhante e, ao mesmo tempo, tão complementar a mim, de maneira tão natural e com tanta amizade, companheirismo e cumplicidade. Amo muito todos vocês.

Agradeço a todos os amigos que me permitiram ter uma válvula de escape deste vórtice, são tantos para tão pouco espaço. Em particular, mas sem ordem particular, agradeço a Flávio, Marta, Ju Alves e Vanessa, minha segunda família por vários anos; Octávio, Patrick, Gui, Cristiano e Robson, companheiros de insanidade musical; Pedro Heisler, Raoni, Bernardo, Vogel, Parada, Caê, João Ricardo, e outros tantos amigos de zoeira ilimitada; Gergely, Hubert e Lyudmila, amigos repentinos em um mundo novo; Bola, 01, Roberto, Lilith, Sheneez, Vanessa, Suelen, Renato, Mislene, Carlos, Sophie, Carol, Diego, Gabriela, Vinícius, Carol Nana, Lucas, Gabriel, Karina, Alex, Ausra, Camila, Karla, Lívia, Letícia, Raíssa, Donatella, e outros tantos amigos de passeios aleatórios pelo Rio e afora.

Agradeço aos professores Edmundo e Rosa, que me permitiram ingressar em seu grupo de pesquisa mais de dez anos atrás, e que sempre serviram de inspiração durante o meu crescimento no meio acadêmico. Agradeço também aos meus co-autores: o professor Daniel, por todos os conselhos e ensinamentos durante a elaboração desta tese, e os professores Valmir e Matthias, pelas colaborações em cada uma de suas partes. Por fim, agradeço ao professor Don, por me permitir fazer parte de seu grupo, mesmo que brevemente, e ao professor Rezende, com quem aprendi muito, mesmo em relativamente pouco tempo de convivência. Inicio minha carreira acadêmica sobre ombros de gigantes.

Agradeço à Carol, por tantos e tantos momentos em que esteve presente durante a montanha-russa destes anos, pelos conselhos durante tantos altos e baixos que muitas vezes pareciam simultâneos. Suas palavras de sabedoria tão bem colocadas

em ocasiões tão necessárias são apenas a ponta do iceberg de todas as diferentes maneira que você ilumina cada um dos alunos que dedica parte de suas vidas para o LAND.

Agradeço aos amigos que fiz durante todos estes anos de UFRJ e de LAND, e que estiveram comigo na fronteira às vezes indissociável entre o profissional e o pessoal. Em particular, mas novamente sem order particular, agradeço a Bernardo, Cadu, Cássio, Couto, Papucaia e Viviane, amigos diários no princípio de nossas trajetórias, a Bernardo, Alysson, GD, Fabrício, Luiz, Gaspare, Guilherme Domingues, Raphael, Carlos, Joanna e Veríssimo, colegas de LAND pelas inúmeras horas de convivência, trocas de ideias e muito trabalho duro; a Ehsan, Brunella, Lucas, Vincent, Julien, Christina, Sebastian, Young-Jun, Emti, Marc e Mohamed, pelo acolhimento durante minha breve permanência no EPFL.

Agradeço a todos os professores, mestres com quem tive oportunidade de absorver nem que fosse um pouco de conhecimento. Em particular, agradeço aos professores Daniel Sadoc, Franklin Marquezino, Maria Eulália Vares e Augusto Teixeira, por terem aceito o convite para compor a banca examinadora desta tese; e aos professores da UFRJ e do IMPA, Jayme Szwarcfiter, Celina Figueiredo, Felipe França, Marta Mattoso, Cláudio Esperança, Henrique Cukierman, Luís Henrique Costa, Mitre Dourado, Vinícius Gusmão e Roberto Imbuzeiro, pelo conhecimento que adquiri em disciplinas.

Resumo da Tese apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Doutor em Ciências (D.Sc.)

DOIS PROBLEMAS SOBRE A RELAÇÃO ENTRE ESTRUTURA E IDENTIDADE EM REDES

Jefferson Elbert Simões

Setembro/2016

Orientador: Daniel Ratton Figueiredo

Programa: Engenharia de Sistemas e Computação

Diversos fenômenos do mundo real podem ser modelados através de redes (grafos) — conjuntos de objetos (vértices) entre os quais codificamos relacionamentos (arestas) de alguma natureza. Redes como a rede neuronal humana ou a rede de amizades do Facebook possuem um papel fundamental no desenrolar de diversos processos. Tendo em vista a quantidade cada vez maior de dados disponíveis sobre suas estruturas, é um dos grandes desafios da atualidade compreender a relação entre identidades de objetos nestas redes e as estruturas que os conectam, seja visando realizar sua identificação precisa ou garantir seu anonimato.

Esta tese aborda dois problemas situados neste contexto. O primeiro problema trata do conceito de simetria como fonte de identidade estrutural para vértices em uma rede e sua relação com estruturas locais em torno destes vértices. Para estudar esta relação, propomos o conceito de *simetria local* utilizando uma modelagem hierárquica baseada em vizinhanças em torno de vértices. Aplicamos esta abordagem ao modelo Erdős-Rényi de grafos aleatórios e provamos a existência de regimes assintóticos de simetria local e de assimetria local, com a transição entre estes regimes ocorrendo em graus médios muito superiores à simetria tradicional.

O segundo problema é conhecido como *casamento de múltiplas redes* e trata da recuperação de um emparelhamento oculto entre vértices de múltiplas redes distintas utilizando informação contida na correlação estrutural entre estas redes. Propomos uma representação matemática deste problema baseada no conceito de hiperpermutação, e apresentamos um modelo para múltiplos grafos de Erdős-Rényi com estruturas correlacionadas. Introduzimos três métricas distintas de descasamento estrutural de hiperpermutações e derivamos diversos resultados sobre o comportamento estatístico destas métricas para o modelo proposto.

Abstract of Thesis presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Doctor of Science (D.Sc.)

TWO PROBLEMS ON THE STRUCTURE-IDENTITY RELATIONSHIP ON NETWORKS

Jefferson Elbert Simões

September/2016

Advisor: Daniel Ratton Figueiredo

Department: Systems Engineering and Computer Science

Several real-world phenomena can be modeled using networks (graphs) — sets of objects (vertices) between which we encode relationships (edges) of some nature. Networks such as the human neural network or the Facebook friendship network have a fundamental role in the development of several processes. Due to the ever increasing amount of data available on their structures, one of the great current challenges is to understand the relationship between identities of objects in these networks and the structures that connect them, whether we aim to precisely perform their identification or assure their anonymity.

This thesis addresses two problems within this context. The first problem deals with the concept of symmetry as source of structural identity for vertices in a network and its relationship with local structures around these vertices. To study this relationship, we propose the concept of *local symmetry* using a hierarchical model based on neighborhoods around vertices. We apply this approach to the Erdős-Rényi model of random graphs and we prove the existence of asymptotic regimes of local symmetry and of local asymmetry, with the transition between these regimes taking place in much higher average degrees than for traditional symmetry.

The second problem is known as *multiple network matching* and deals with the recovery of a hidden matching between vertices of multiple distinct networks using information obtained from the structural correlation between these networks. We propose a mathematical representation of this problem based on the concept of hyperpermutation, and present a model for multiple Erdős-Rényi graphs with correlated structures. We introduce three distinct statistics for structural mismatch of hyperpermutations and we derive several results about their statistical behavior for the proposed model.

Contents

List of Figures

Chapter 1

Introduction

"Dreams are seldom shattered, by a bullet in the dark
Rulers come and rulers go, will our kingdom fall apart?"
(Joakim Brodém, Pär Sundstrom)

Abstractly, we can define a network as a set of actors or objects, among which are identified, exogenous or endogenously, end-to-end ties or relationships of a certain nature. In this context, is it undeniable that our world is structured around numerous networks. Some networks, such as the Internet and the highway mesh, define the way human beings interact, while others, such as the academic collaboration network and the sexual contact network, are a snapshot of our interactions. Neural networks and blood vessel networks foreground mechanisms that define how our body works, while networks of interacting particles explain properties of several substances in the universe.

Even though networks (or graphs, as known in contexts such as mathematics and computer science) have been studied since Euler's "Seven Bridges of Königsberg" and Vandermonde's "knight's tour" problems, and profound results have been known for decades, in the last twenty years a new research field has emerged known as *network science* [1], whose goal is to jointly study and understand all these networks, their similarities and differences, the mechanisms that govern them and the mechanisms that they govern.

One of the main reasons this research field has risen so quickly is the great amount of data produced and made available about these networks. For this reason, it is extremely important to be able to reliably identify the actors in these networks, or avoid identification thereof. For instance:

- The study of research collaboration networks (graphs formed by researchers, who are connected if they have published any work in collaboration) requires us to correctly identify co-authorship. Therefore, we must be able to detect

that a single author has publications under different names, or that two distinct authors with similar names have published articles using the same name. Problems such as these have been studied in the context of *name disambiguation* [2];

- The evaluation of properties of social networks requires the understanding of people's features and, in some cases, there is a strong private character in this information. In order to make data about these networks available, we must preserve the anonymity of the people that it comprises, leading to the study of *social network anonymization* [3].

The relationship between the structure of a network and the identity of its vertices has been studied from several distinct points of view and with several different names, constituting a broad field of knowledge which we call *structural identity in networks*. Its ancestry traces mainly to works on *structural equivalence* from mathematical sociology in the 1970's, however, recent work has led to very interesting advances and have brought the field into an intersection with computer science and engineering. Moreover, since its object of interest is neither the abstract notion of network nor a specific kind of real network, but the structure-identity relationship as observed in general real networks, structural identity can as well be thought of as being a subfield of network science as well, despite its overall focus on social networks where the more immediate application of the concept of identity sparks more interest in general.

This thesis dissertates on two distinct problems in the field of structural identity in networks. The first problem is the application of the concept of symmetry as a tool for analyzing a network's structure, and the second problem aims at recovering a hidden matching between the identities of objects in distinct networks by leveraging their structural similarity. We will refer to these problems as *symmetry in graphs* and *graph matching*, respectively. As we will argue further, most of the work in the field of structural identity follows either an analytical or a practical approach, and these two problems are an example of this duality, as each one is located in a different branch.

Our contributions can be summarized as follows. On symmetry in graphs, we consider the relationship between vertex identities and localities by proposing a definition of *local symmetry*, based on the structural similarity of neighborhoods around each vertex. Our definition naturally induces a hierarchy of symmetries, which progressively uses more information for classifying vertices, ultimately culminating in the traditional, automorphism-based symmetry, which we call *global symmetry* in the context of this work. For the most restricted case of local symmetry, we identify regimes of asymptotic local symmetry and asymptotic local asymmetry for the

Erdős-Rényi random graph. We find that, relative to global symmetry, asymptotic local symmetry persists for a much broader range of average degrees, even though asymptotic local asymmetry still emerges eventually. We also present numerical results obtained via simulation in order to provide experimental confirmation for these findings.

On the problem of graph matching, we focus on a version of this problem called the *multiple graph matching*. This problem is a generalization of the traditional network matching problem, as we move from matching the structures of two networks to an arbitrary number of networks. Our goal is to understand the fundamental feasibility limits of this problem in the absence of any context-dependent information, that is, based only on the information given by the structural similarity of the networks at hand. For this reason, our approach is based on mathematical modelling and analysis, in contrast with the applied approach taken by most works on this problem. We propose the concept of *hyperpermutations* as a representation of a matching between multiple sets, and introduce three distinct statistics for structural mismatch between multiple graphs, with the goal to reveal the underlying, correct matching. Furthermore, we introduce a model for structurally-correlated random graphs, called the $G_k(n, p, s)$ model, and derive a number of preliminary results regarding the statistical behavior of our mismatch statistics under this model.

The remainder of this thesis is structured as follows. In Chapter 2, we present several preliminary concepts on which both halves of our work are based, including several concepts in graph theory and random graph models. Chapters 3 and 4 deal with the problem of symmetry in graphs: in Chapter 3 we present a brief review of the field, including fundamental concepts, known applications and open problems, while in Chapter 4 we present our proposals and contributions. A similar pattern is followed in Chapters 5 and 6 as we deal with the problem of network matching: we present an overview of recent approaches in this field in Chapter 5, including existing mathematical models, common applications and known results, while our proposals and contributions are given in Chapter 6. Finally, we dedicate Chapter 7 for a summary of our discussion and possibilities of future work.

Chapter 2

Preliminary concepts

"There is a bond between us
Even if it's frayed it is unbreakable"
(Mikael Åkerfeldt)

2.1 Graphs

Mathematically, a graph G is a pair of sets (V, E), V being its set of *vertices* and $E \subseteq \binom{V}{2}$ its set of *edges*[1] [4, p.2] (we will assume, unless stated otherwise, that V is a finite set). This mathematical tool allows us to model real networks by the simultaneous representation of a set of objects (corresponding to vertices) and a well-defined set of pairwise relationships among these objects (corresponding to the presence or absence of edges). Such relationships can have a physical nature. For instance, a road network has a set of geographical points as its vertex set, with edges corresponding to roads directly connecting these points, and the simplest graph representation of the Internet takes all routers as the vertex set and pairwise physical connections between them, usually by means of optical fiber lines, as its edge set. The human neural network, formed by our set of neurons, takes its connections from the synapses that allow electrical signals to flow from one neuron to another, and is thus another, more microscopic example of a physical network.

However, several important networks have an abstract connotation. For instance, a friendship network is a social network that captures the existence of friendship relationships (edges) between people (vertices) within a given context, and a research collaboration network represents the collaborations (edges) between the researchers (vertices) of an institution or a knowledge field. Likewise, the protein-protein interaction (PPI) network of a given species has a set of proteins, found in this species'

[1] We denote by $\binom{V}{k}$ the set of subsets of V with cardinality k. In particular, $\binom{V}{2}$ denotes the set of unordered pairs of elements in V.

cells, as its vertex set, with the existence of edges corresponding to pairwise interactions between these proteins. The relationships between objects in these cases (friendship, collaboration, and interaction) are abstract per se, even though they still stand out physically in multiple ways (such as Facebook friend status, co-authorship of papers, and electrostatic forces).

Two vertices $u, v \in V$ are said to be *adjacent*, or neighbors, if $\{u, v\} \in E$, and we write $u \sim_G v$; otherwise, we write $u \not\sim_G v$ (we will omit the index $_G$ whenever the graph at hand can be inferred from the context). The *degree* of a vertex v ($d(v)$) is the number of vertices that are adjacent to v, and the *distance* between two vertices u and v ($d(u, v)$) is the smallest k such that there is a sequence of vertices $u = w_0, w_1, w_2, \ldots, w_k = v$ sequentially adjacent (that is, with $w_{i-1} \sim_G w_i$ for $1 \leq i \leq k$). Naturally, the distance between adjacent vertices is exactly 1.

Given a graph $G = (V, E)$ and a vertex set $S \subseteq V$, we define $\mathcal{N}_G(S)$, the *open neighborhood* of S in G:

$$\mathcal{N}_G(S) = \{v \in V \ : \ d(v, S) \leq 1\},$$

and the *closed neighborhood* of S in G:

$$\mathcal{N}_G[S] = G[\mathcal{N}_G(S)],$$

where, for every set $A \subseteq V$ of vertices, $d(v, A)$ is the smallest distance between v and some vertex of A, and $G[A]$ is the subgraph induced by the vertices of A[2]. Naturally, $\mathcal{N}_G(\{v\})$ (or, by simplicity, $\mathcal{N}_G(v)$) is the set that comprises v and its neighbors, and $\mathcal{N}_G[\{v\}]$ (or $\mathcal{N}_G[v]$) is the subgraph induced by v and its neighbors. Again, we will omit the index $_G$ whenever possible.

This definition of neighborhood is traditional in the field of graph theory[3], and we can extend it to include not only vertices at distance 1, but at distance k. We achieve this goal easily by defining the *open k-neighborhood* of S in G:

$$\mathcal{N}^k_G(S) = \{v \in V \ : \ d(v, S) \leq k\}.$$

Note that $\mathcal{N}^0(S) = S$, $\mathcal{N}^1(S) = \mathcal{N}(S)$ and, recursively, $\mathcal{N}^k(S) = \mathcal{N}(\mathcal{N}^{k-1}(S))$. We define the *closed k-neighborhood* of S in G similarly:

$$\mathcal{N}^k_G[S] = G[\mathcal{N}^k_G(S)].$$

[2]A subgraph induced by a set of vertices comprises the vertices of this set and all edges between these vertices.

[3]Actually, the convention in graph theory is to define $\mathcal{N}_G(S)$ as the subgraph induced by the neighbors of S, disregarding vertices of S itself. We choose to use this alternate definition to ease the generalization to higher-order neighborhoods.

We illustrate these definitions in Figure 2.1. Note that $\mathcal{N}^k(S) \subseteq \mathcal{N}^{k+1}(S)$ and that $\mathcal{N}^k[S]$ is an induced subgraph of $\mathcal{N}^{k+1}[S]$. We also note that, in contexts such as mathematical sociology, the closed neighborhood of a vertex is known as an *egonet* [5, 6].

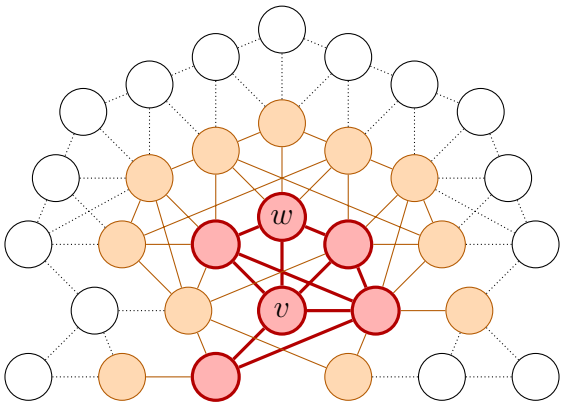


Figure 2.1: Examples of neighborhoods of a vertex in a graph. The subgraph highlighted in red is $\mathcal{N}^1[v] = \mathcal{N}[v]$, and including the orange highlights we obtain $\mathcal{N}^2[v]$. The vertex sets of these subgraphs are, respectively, $\mathcal{N}^1(v) = \mathcal{N}(v)$ and $\mathcal{N}^2(v)$. In this example, for $k \geq 3$, $\mathcal{N}^k[v] = G$.

2.2 Random graphs

Random graph models appeared in the 50's and, more recently, have been acknowledged as a better representation for networks resulting from real-world processes. In these models, the resulting graph is obtained from some probabilistic model, its structure being represented by a probability distribution over a pre-defined set of graphs. The most common form of expressing these models is through a generative process, although some models may be explicitely described by the actual probability distribution they impose over the graph set. In this context, all properties and measures over these graphs become random.

Formally, a random graph model $\mathcal{M}$ is a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where Ω is a set of graphs that can be obtained by the model (samples), $\mathcal{F} \subseteq \mathcal{P}(\Omega)$ is a σ-algebra of subsets of samples (events), and $\mathbb{P}$ is a probability measure over these events. A common and usually harmless abuse of notation is to use the symbol $\mathcal{M}$ that represents the model to also represent a random sample of this model — we will follow this whenever possible. Most models implicitly consider that Ω is the set of graphs $G = (V, E)$ with fixed $|V| = n$, $V = \{1, 2, \ldots, n\}$ by convention whenever necessary, and $\mathcal{F} = \mathcal{P}(\Omega)$, the set of parts of Ω, such that the model is described and determined by the measure $\mathbb{P}$, or by a process that allows to characterize its properties.

In this scenario, properties of the graphs at hand turn to be represented by events in $\mathcal{F}$ and, being $\mathcal{F}$ the biggest possible σ-algebra, every property is measurable and has a probability of being observed. Similarly, every measurement on these graphs is represented by a random variable (or, more generally, a random element) in this probability space, whose behavior will be described by their distributions. Given the difficulty usually imposed by the model formulation in determining precisely the probability of events and distribution of random variables of interest, it is also commonplace to undertake analysis that are asymptotic with the graph size — that is, results that apply when $n \to \infty$ — over these probabilities, considering that the remaining model parameters are not necessarily constant but, potentially, functions of n.

A particularly frequent approach consists in, given a model $\mathcal{M}$, search for asymptotic regimes in which the probability of observing a certain property Q approches 1 (or, equivalently, the probability of its negation approaches 0). In this case, we will say that, in such regimes, $\mathcal{M}$ exhibits Q a.a.s. (asymptotically almost surely). This term is similar to the expression a.s. (almost surely) used to denote events of properties that occur with probability equal to 1.

Throughout this work, we will use the standard probability theory notation for asymptotic behavior of functions: given two positive functions $f(n), g(n)$, we say that $f = o(g)$ (or $g = \omega(f)$) if, for every constant $c > 0$, it is true that $f(n) < cg(n)$ for large enough n. We will use the terms $o(\cdot)$ and $\omega(\cdot)$ also to refer to some function satisfying these properties. Also, we will say that $f = O(g)$ (or $g = \Omega(f)$) if, for some constant $c > 0$, it is true that $f(n) \leq cg(n)$.[4] The latter notation is more common to the computer science than probability theory. Note that, according to these definitions, a model $\mathcal{M}$ exhibits property Q a.a.s if and only if $\mathbb{P}[\mathcal{M} \text{ exhibits Q}] = 1 - o(1)$.

With the emergence of more intense research in network science in the last two decades, it was necessary to formulate models to capture features often observed in real-world networks. To exemplify with two very well-known and studied models, the Barabási-Albert model [7] proposes a random generative process, based on a cumulative advantage mechanism, whose resulting graphs will exhibit, with high probability, heavy-tailed degree distribution; that is, they will have vertices whose degree exceeds the average degree in the graph by several orders of magnitude. This property is observed in several networks distinct by nature, such as the network of web pages and collaboration networks between researchers. Another property frequently observed is known as "small world effect": networks with such properties — such as gene regulatory networks, the network of autonomous systems on the Internet and, again, the Web network — exhibit vertex-to-vertex distances typically

[4]Note that $f = o(g)$ implies $f = O(g)$ and, analogously, $f = \omega(g)$ implies $f = \Omega(g)$.

much smaller than the size of network, despite being sparse networks. The Watts-Strogatz model [8] captures this phenomenon applying a few random modifications in a ring-shaped regular lattice graph.

Older than these, the two most studied random graph models are ambiguously called Erdős-Rényi model (or ER model). Both were proposed at the 50's, one by Paul Erdős and Alfréd Rényi, and the other one by Edgar Gilbert. In the first model, denoted by $G(n, m)$, a graph is picked uniformly at random among the graphs on n vertices and m edges [9]. In the second one, denoted by $G(n, p)$, a graph on n vertices is constructed by adding an edge between each pair of vertices with probability p and independently of the remaining edges [10].

The relative simplicity of these descriptions allows for the study of several properties on both models, conferring them a very important role in network mathematical modelling. For instance, a property that applies to "almost all graphs" can be formally defined as being a property that is exhibited by the $G(n, p = {}^1/_2)$ model (in which all graph are equiprobable) a.a.s. [11, p.43]. Erdős used the $G(n, p)$ and $G(n, m)$ models to demonstrate theorems about the existence of graphs with certain properties, making room in graph theory for the probabilistic method [12, p.1], a technique for proving deterministic results using probabilistic models and arguments, strongly used in areas like Ramsey theory and number theory. Several works relied on the $G(n, p)$ model for formulating mean field approximations for graph percolation models [13].

Chapter 3

Symmetry in graphs

"I am cast out and I am not like you
Find my way on through the haze
I am liquefied in a strange brew"
(Mikael Åkerfeldt)

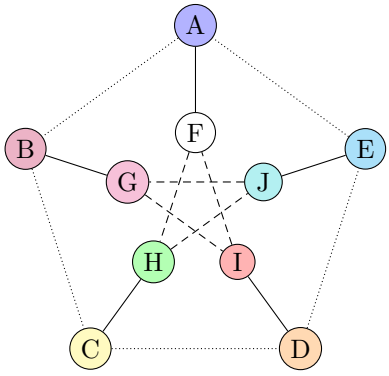
In mathematics and physics, the abstract concept of symmetry usually refers to the invariance of a given structure when subjected to a class of transformations. Each structure of mathematics is deemed symmetric by its own properties: symmetric matrices are invariant with respect to the flip operation over its diagonal; a symmetric probability distribution over real numbers is invariant with respect to a horizontal flip of its graph around some given value of its domain; fractals such as the Mandelbrot set and the Sierpinski triangle are scale-symmetric, that is, invariant to operations of zoom-in and zoom-out; and geometric objects and manifolds can portray a number of distinct symmetry properties, such as rotational, translational and helical symmetry.

In the context of graphs, this concept is traditionally embodied by the notions of *isomorphism* and *automorphism*. Two graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ are said to be *isomorphic* if there is a bijective function $f : V_1 \to V_2$, called an *isomorphism* between G_1 and G_2, that precisely maps the edges of G_1 into edges of G_2, thus satisfying the following property:

$$\forall\, u, v \in V_1 \;:\; u \sim_{G_1} v \iff f(u) \sim_{G_2} f(v)$$

An isomorphism between G and G itself is called an *automorphism* of G [4, p.14]. Figure 3.1 visually illustrates these definitions on the well-known Petersen graph.

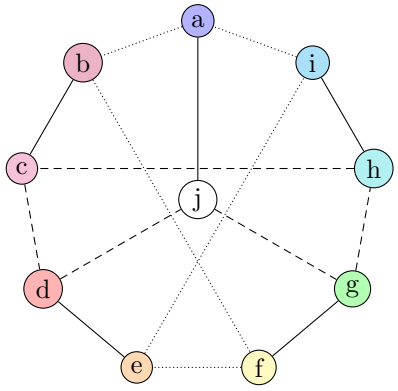
Not every pair of graphs exhibits an isomorphism, but every graph has at least one automorphism, namely, the identity function $\mathbb{I}_V$ that maps every vertex on itself — such automorphism is called *trivial*, and a graph having only the trivial automorphism is said to be *asymmetric* (the graph G_1 presented in Figure 3.1d,



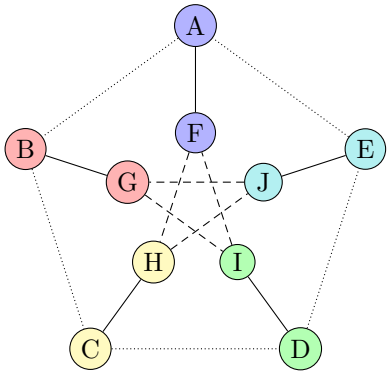
(a) G_1

v	$f(v)$
A	a
B	b
C	f
D	e
E	i
F	j
G	c
H	g
I	d
J	h

(b) Isomorphism between G_1 and G_2



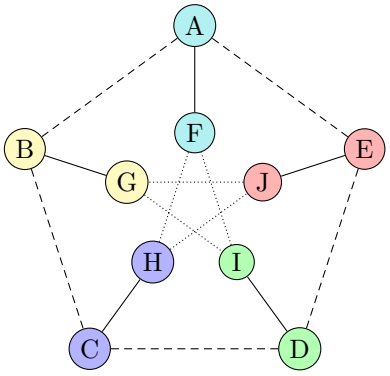
(c) G_2



(d) G_1

v	$f'(v)$
A	H
B	J
C	G
D	I
E	F
F	C
G	E
H	B
I	D
J	A

(e) Automorphism of G_1



(f) G_1

Figure 3.1: Examples of isomorphism between two graphs G_1 and G_2 (figures a–c) and automorphism of G_1 (figure d–f). In all diagrams, circles and lines connecting them represent vertices and edges, respectively. Vertices are colored according to the selected isomorphism/automorphism. Tracing patterns on the edges were chosen as a visual guide.

for instance, is not asymmetric, as shown by the nontrivial automorphism f'). An open question in theoretical computer science is to determine the computational complexity of deciding whether a given graph has a nontrivial automorphism, or whether two given graphs are isomorphic. It is known that both problems belong to the class NP [14], which means it is possible to computationally check the existence of an isomorphism between two graphs, or a nontrivial automorphism of a graph, in polynomial time as a function of the size of the graph(s) (using the iso/automorphism itself as a certificate) [15]. However, it is still unknown whether these problems can be solved in polynomial time, or whether they are at least as difficult to be solved as the most difficult problems in NP — in other words, they are not known to be either in the class P or NP-complete.

The problem of graph isomorphism, in particular, has been deeply studied in recent decades: in an effort to understand its computational complexity, it has become the basis of a new complexity class called the GI class [16], defined as the class of problems equivalent in polynomial time to the graph isomorphism problem. The fastest known theoretical algorithm, with time complexity $2^{O(\sqrt{n \log n})}$, is due to Luks [17] combined with a result by Zemlyachenko et al. [18], though a quasipolynomial time algorithm proposed by Babai [19] is currently under a proof-checking process. Applied results include a number of proposed heuristic-based algorithms [20–23], several of which have empirically shown reasonable performance [24].

In graph theory and discrete mathematics, the property "not being asymmetric" is the most general among those that intend to capture symmetry patterns in a graph. Other traditional definitions impose conditions even stronger than the mere existence of a non-trivial automorphism, and usually imply a strong equivalence between all vertices of the graph at hand. For instance, for a graph G to be *vertex-transitive* [4, p.14], it is necessary that, for every pair of vertices u, v in V, there is an automorphism f such that $f(u) = v$ (all vertices can be mapped to one another by some automorphism). Even more restrictive is the definition of a *symmetric*[1] graph G [25, p.104]: it requires that, for every pair of edges $\{u, v\}$ e $\{w, x\}$, it is possible to map u to w and v to x with the same automorphism f (that is, $f(u) = w$ e $f(v) = x$).

These properties, though widely studied and observed in graph models originated in several fields (such as the Ising model in two dimensions [26], in statistical mechanics), require an equivalence too strong between its vertices, which makes them very difficult to find in real-world networks. For instance, for a road network to be well represented by a transitive graph, it would be necessary that every intersection had the same number of exits, which is an unrealistic assumption for this scenario. On the other hand, the concepts of automorphism and isomorphism have been successfully used in practical problems such as network privacy and anonymization [27] and computer vision [28]. Let us explore these concepts a bit further.

The isomorphism relation between pairs of graphs is reflexive (due to the trivial automorphism), symmetric (due to the existence of the inverse of any isomorphism) and transitive (by the composition of isomorphisms), therefore it is an equivalence relation on the set of all graphs. The most natural interpretation to this fact is that graphs can be reduced, through this relation, to equivalence classes defined only by the structure of the graphs belonging to each class, no matter how arbitrary and distinct are the elements of the sets of vertices of these graphs. This interpretation is implicit in traditional graph theory expressions such as uniqueness of a graph *up to*

[1] Note how, in the traditional definition, the properties "is symmetric" e "is not asymmetric" are not complementary.

isomorphism satisfying certain structural properties, so that the nature or identity of the vertices is irrelevant and the graph is reduced, for the purpose of analysis, to its edge pattern.

Similarly, the set of automorphisms of a graph, usually denoted $\mathrm{Aut}(G)$, possesses an algebraic group [4, p.18] structure when equipped with the operation of function composition. Moreover, this group structure induces an equivalence relation between the vertices of this graph: the relation "can be mapped by an automorphism", which too is reflexive, symmetric and transitive by the same arguments as before. The interpretation now is that the vertices can be grouped according to their placement in the graph structure, such that vertices in the same class are "structurally indistinguishable", at least without additional information about the identities of (potentially all) the remaining vertices.

This interpretation leads us to the following definitions:

Definition 3.1. *Given a graph $G = (V, E)$, two vertices $v_1, v_2 \in V$ are* globally symmetric *if there is an automorphism f of G such that $f(v_1) = v_2$.*

Definition 3.2. *Let $G = (V, E)$ be a graph. Then G is said to be* globally symmetric *if there are $u, v \in V$ distinct and globally symmetric.*

Remark 3.3. *G is globally symmetric if and only if it has at least one non-trivial automorphism.*

Throughout this thesis, we will use the term "globally asymmetric" both for pairs of vertices and for graphs that are not globally symmetric, and use these definitions of global symmetry as the standard portrayal of the concept of symmetry in graphs. Alternative definitions, specifically in the literature of probability theory, have employed the simpler term *symmetric graph* for graphs satisfying Remark 3.3 [29], but this conflicts with at least two known graph-theoretic definitions for the term "symmetric graph" [25, p.104] [30], including the one already described. Thus, introducing a new term in these definitions may help avoid ambiguities.

Global symmetry has been extensively studied in random graph models. It is known [29] [11, p.230] that, for the $G(n, m)$ model, a globally asymmetric graph is obtained a.a.s. if and only if $2m/\binom{n}{2} \geq \log n + \omega(1)$ and $n - 1 - 2m/\binom{n}{2} \geq \log n + \omega(1)$ — that is, when the model exhibits average degree at least slightly larger than $\log n$ and at most slightly smaller than $n - 1 - \log n$. For $G(n, p)$, an analogous but slightly weaker result is known: the $G(n, p)$ random graph is globally asymmetric a.a.s. if $p \in [\log n/n, 1 - \log n/n]$ [31], and it is globally symmetric if $p = o(\log n/n)$ (due to the existence of isolated vertices a.a.s. [32]) or $1 - p = o(\log n/n)$ (due to the existence of universal vertices a.a.s.).

One of the greatest difficulties in studying symmetry in graphs, which is an open problem to the best of our knowledge and is not addressed in this thesis,

lies in quantifying how much a graph is "almost symmetric". The usefulness of such quantification comes from the fact that, in many real-world networks, one can identify intuitively equivalent vertices that the definition of global symmetry falls short in capturing. For instance, consider the network in Figure 3.2. Intuitively, the vertices u and v can be seen as small "hubs", who share almost equally the role of connecting the remaining, "peripheral" vertices, which are also intuitively equivalent to each other. However, since u and v have different degrees, no automorphism is able to map them onto one another, or map a peripheral neighbor of u onto a peripheral neighbor of v. Thus, the equivalence classes of vertices found on this graph do not capture this intuitive notion of symmetry.

Note that this approach to symmetry as a binary property, which graphs simply might or might not have, also implies that small changes to a graph's structure, such as the insertion or removal of a single edge, might be able to suddenly make symmetry appear or disappear, which also conflicts with the intuitive concept of an object's symmetry as a robust property, dependent on its own structure rather than individual pieces.

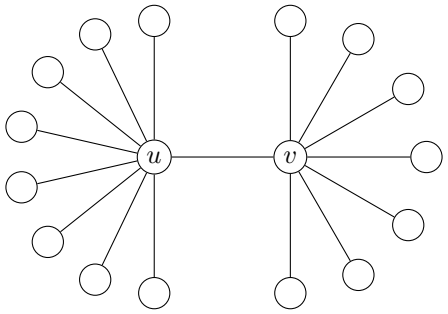


Figure 3.2: Example of "almost symmetric" graph. Vertices u and v connect their neighbors to the remainder of the graph.

Several measurements of "almost-symmetry" have been proposed in the literature, usually by counting the number of modifications to be imposed to the graph at hand so that is acquires a non-trivial automorphism. Such modifications can be restricted to addition and removal of edges [29] or include more convoluted operations, such as edge contractions [33]. However, no proposal is known to have achieved great acceptance, because of both the computational complexity of calculating these measurements and the absence of knowledge about the relationship between them and the performance of applications or real-world processes on networks that require their symmetry (of absence thereof).

A second difficulty, which we address in this thesis and that has barely been analyzed in the literature, lies in capturing an intuitive conceptualization of symmetry between vertices representing a similar positioning with respect to their respective structures, even if these vertices might be globally distinguishable. The graph in

Figure 3.3 illustrates this idea. In this figure, we highlight two isomorphic induced subgraphs. However, since both these subgraphs are part of a larger graph, the existence of an automorphism that reflects such symmetry depends on the edge pattern of the remainder of the graph, and in this particular example the desired automorphism does not exist. In other words, even though these subgraphs are intuitively symmetric, since they are connected differently to the graph, they are not globally symmetric.

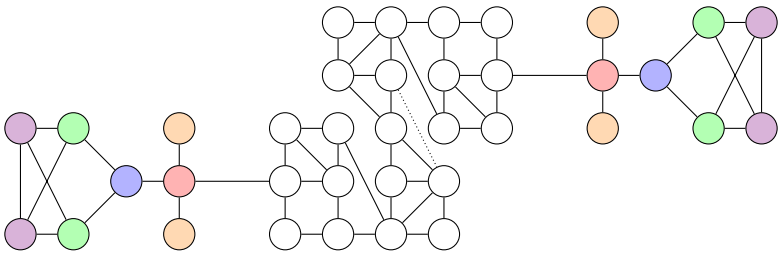


Figure 3.3: Example of equivalence between local structures in a graph.

Note that, regardless of the structure of the remainder of the graph, and despite the fact that the intuitive equivalence between the two local structures continues to exist, the two aforementioned difficulties can happen simultaneously. For instance, in the example of Figure 3.3, by removing the single edge identified by a dashed line and located at least at distance 5 from the highlighted subgraphs, the graph acquires an automorphism that maps these subgraphs precisely. Hence, the existence of this automorphism — a global mapping satisfying local restrictions — is also sensitive to changes that seem unpretentious and unrelated to the local structures of interest.

Chapter 4

Local symmetry

"As you're facing the path that divides
I will always be here by your side"
(John Petrucci)

4.1 Definitions

To proceed more precisely, we must first delineate which kinds of "locality" we are interested in analyzing. For the purposes of this thesis, we will use closed k-neighborhoods around single vertices as a proxy for "locality". This motivates the following definition:

Definition 4.1. *Given a graph $G = (V, E)$, two vertices $v_1, v_2 \in V$ are* k-locally symmetric *if there is an isomorphism f between $\mathcal{N}^k[v_1]$ and $\mathcal{N}^k[v_2]$ such that $f(v_1) = v_2$.*

Therefore, two vertices v_1 and v_2 are k-locally symmetric if the k-th order local structures in which v_1 and v_2 are located are equivalent, with v_1 and v_2 equivalently located in these structures. Note that $\mathcal{N}^k[v_1]$ and $\mathcal{N}^k[v_2]$ are not necessarily disjoint. If $k = 1$, for simplicity, we will say that the vertices are *locally symmetric*.

One interesting feature of k-local symmetry is that it naturally leads to the construction of a symmetry hierarchy, which includes global symmetry as the most restrictive one, as evinced by the following result. Let the diameter [4, p.71] of a graph G ($\mathrm{diam}(G)$) be the biggest distance between any pair of its vertices.

Proposition 4.2. *Let v_1 and v_2 be vertices of $G = (V, E)$, and let $k \in \mathbb{N}$. Then the following statements hold:*

1. *If v_1 and v_2 are $(k+1)$-locally symmetric, then v_1 and v_2 are k-locally symmetric;*

2. *If $k \geq \text{diam}(G)$, then v_1 and v_2 are k-locally symmetric if and only if v_1 and v_2 are globally symmetric.*

Proof. If v_1 and v_2 are $(k+1)$-locally symmetric, then there is an isomorphism f between $\mathcal{N}^{k+1}[v_1]$ and $\mathcal{N}^{k+1}[v_2]$ with $f(v_1) = v_2$. Since isomorphisms preserve distances, for any $d \in \mathbb{N}$, two vertices u and v of $\mathcal{N}^{k+1}[v_1]$ are at distance d if and only if $f(u)$ and $f(v)$ are at distance d in $\mathcal{N}^{k+1}[v_2]$. Therefore, for every vertex w in $\mathcal{N}^{k+1}[v_1]$, it is true that $w \in \mathcal{N}^k[v_1] \iff f(w) \in \mathcal{N}^k[v_2]$. This means that $f|_{\mathcal{N}^k[v_1]}$ is a bijection between $\mathcal{N}^k(v_1)$ and $\mathcal{N}^k(v_2)$. Note that, since f also preserves edges, so does $f|_{\mathcal{N}^k[v_1]}$, thus $f|_{\mathcal{N}^k[v_1]}$ is an isomorphism between $\mathcal{N}^k[v_1]$ and $\mathcal{N}^k[v_2]$. Since $f|_{\mathcal{N}^k[v_1]}(v_1) = f(v_1) = v_2$, it follows that v_1 and v_2 are, by definition, k-local symmetric.

This proves the first statement. The second statement follows from the fact that, if $k \geq \text{diam}(G)$, then any two vertices of G are at distance k or smaller from each other. Therefore, $\mathcal{N}^k(v) = V$ and $\mathcal{N}^k[v] = G$, and the definitions of k-local symmetry and global symmetry are equivalent. □

Finally, we define local symmetry in graphs:

Definition 4.3. *Let $G = (V, E)$ be a graph. Then G is* k-locally symmetric *if there are $u, v \in V$ distinct and k-locally symmetric.*

Note that this definition for k-local symmetry is analogous to Definition 3.2 for global symmetry. Therefore, our definitions of k-local symmetry — for two vertices and for a single graph — are mutually consistent in the same fashion as those of global symmetry. This definition also implies a symmetry hierarchy similar to Proposition 4.2:

Proposition 4.4. *Let $G = (V, E)$ be a graph, and let $k \in \mathbb{N}$. Then the following statements hold:*

1. *If G is $(k+1)$-locally symmetric, then G is k-locally symmetric;*

2. *If $k \geq \text{diam}(G)$, then G is k-locally symmetric if and only if G is globally symmetric.*

Given our interest in understanding symmetry in random graph models, it is natural that we start our analysis with the $G(n, p)$ model, which is more analytically tractable than most other models. In particular, we would like to determine if there is any fundamental difference between the emergence of local symmetry and global symmetry in this model. Our results, which we present in the following sections, show that such fundamental difference indeed exists, as local symmetry persists to much higher average degrees in the $G(n, p)$ random graph than global symmetry.

4.2 Symmetry regime

Our first result is the identification of a 1-local symmetry regime for $G(n,p)$:

Theorem 4.5. *A $G(n,p)$ random graph, with $p = o(n^{-2/3})$, is locally symmetric a.a.s.*

Let us proceed with some terminology before the proof of this statement. Recall the definition of closed 1-neighborhood of a vertex. We call this vertex the *center* of this subgraph, with all edges between the center and other vertices said to be *core edges*, and all remaining edges termed *peripheral edges*.

The idea behind the proof is that peripheral edges always close triangles, but in this regime, $G(n,p)$ does not have too many triangles, so the closed neighborhood of many vertices are simple stars. Two such vertices will have isomorphic closed neighborhoods simply by having the same degree. Now, since the degrees in $G(n,p)$ concentrate heavily around their mean, the degree sequence of $G(n,p)$ has small support, which means there must be at least two vertices with the same degree among those with star neighborhoods, which proves the result.

To execute the proof, we will need two auxiliary results:

Lemma 4.6. *Let $G = (V,E)$ be a $G(n,p)$ random graph. If $p = \omega(\log n/n)$, then for any fixed $\delta \in (0,1)$, the degree of all vertices of G are within the range $(n-1)p(1\pm\delta)$, a.a.s.*

Proof. Let d_v be the degree of vertex v. We know that $d_v \overset{d}{\sim} \mathrm{Bin}(n-1,p)$. By the Chernoff bound, for any $\delta \in (0,1)$:

$$\mathbb{P}[d_v \notin (n-1)p(1\pm\delta)] \leq 2e^{-(n-1)p\delta^2/3}.$$

The union bound implies:

$$\mathbb{P}[\exists\ v\ :\ d_v \notin (n-1)p(1\pm\delta)] \leq 2ne^{-(n-1)p\delta^2/3}$$

and, since $np = \omega(\log n)$ by hypothesis, the right-hand side is $2ne^{-\omega(\log n)} = 2no(1/n) = o(1)$. □

Lemma 4.7. *Let $G = (V,E)$ be a $G(n,p)$ random graph, and let T be the number of triangles in G. Then, $\mathbb{E}[T] = \binom{n}{3}p^3$ and, if $p = \omega(1/n)$, $\mathbb{P}[|T - \mathbb{E}[T]| < c\cdot\mathbb{E}[T]] \to 1$ for any fixed $c > 0$.*

Proof. We denote by $\binom{V}{3}$ the set of unordered triples of vertices, and for each triple $t = (i,j,k)$, we define the event $\{t \text{ is } \Delta\} = \{(i,j),(i,k),(j,k) \in E\}$. Then:

$$T = \sum_{t\in\binom{V}{3}} \mathbb{1}_{\{t \text{ is } \Delta\}}.$$

It easily follows from linearity of expectation and independence of edges that $\mathbb{E}[T] = \sum_{t \in \binom{V}{3}} \mathbb{P}[t \text{ is } \Delta] = \binom{n}{3} p^3$.

We also need to estimate the variance of T, which we denote by $\mathbb{V}[T]$. For such, we need an expression for its second moment:

$$\begin{aligned}
\mathbb{E}[T^2] &= \mathbb{E}\left[\left(\sum_{t \in \binom{V}{3}} \mathbb{1}_{\{t \text{ is } \Delta\}}\right)\left(\sum_{t \in \binom{V}{3}} \mathbb{1}_{\{t \text{ is } \Delta\}}\right)\right] \\
&= \sum_{t,t' \in \binom{V}{3}} \mathbb{E}[\mathbb{1}_{\{t \text{ is } \Delta\}} \mathbb{1}_{\{t' \text{ is } \Delta\}}] \\
&= \sum_{t,t' \in \binom{V}{3}} \mathbb{P}[t \text{ is } \Delta, t' \text{ is } \Delta]
\end{aligned}$$

This summation can be broken into four pieces, based on the relationship between the two triples of vertices, t and t':

No common vertices All edges of t are independent of all edges of t', so $\mathbb{P}[t \text{ is } \Delta, t' \text{ is } \Delta] = p^6$;

One common vertex Again, edges of t are independent of edges of t', and $\mathbb{P}[t \text{ is } \Delta, t' \text{ is } \Delta] = p^6$;

Two common vertices t and t' share the edge between common vertices, comprising a total of 5 edges — thus, $\mathbb{P}[t \text{ is } \Delta, t' \text{ is } \Delta] = p^5$;

Three common vertices In this case, $t = t'$ and $\mathbb{P}[t \text{ is } \Delta, t' \text{ is } \Delta] = p^3$.

We must also count how many triples fit into each of these cases:

No common vertices $\binom{n}{3}\binom{n-3}{3}$ triples;

One common vertex $\binom{n}{3} \cdot 3\binom{n-3}{2}$ triples;

Two common vertices $\binom{n}{3} \cdot 3(n-3)$ triples;

Three common vertices $\binom{n}{3}$ triples.

We can now calculate the second moment of T:

$$\begin{aligned}
\mathbb{E}[T^2] &= \sum_{t,t' \in \binom{V}{3}} \mathbb{P}[t \text{ is } \Delta, t' \text{ is } \Delta] \\
&= \binom{n}{3} p^3 + 3\binom{n}{3}(n-3)p^5 + 3\binom{n}{3}\binom{n-3}{2}p^6 + \binom{n}{3}\binom{n-3}{3}p^6 \\
&= \binom{n}{3} p^3 \left[1 + 3np^2 - 9p^2 + \frac{3}{2}n^2p^3 - \frac{3}{2}7np^3 + \frac{3}{2}12p^3 \right. \\
&\qquad \left. + \frac{1}{6}n^3p^3 - \frac{1}{6}12n^2p^3 + \frac{1}{6}47np^3 - \frac{1}{6}60p^3\right].
\end{aligned}$$

We can also obtain an expression for $\mathbb{E}[T]^2$:

$$\begin{aligned}
\mathbb{E}[T]^2 &= \left(\binom{n}{3}p^3\right)^2 \\
&= \binom{n}{3} p^3 \left[\frac{1}{6}n^3p^3 - \frac{1}{6}3n^2p^3 + \frac{1}{6}2np^3\right].
\end{aligned}$$

Combining these expressions yields an expression for $\mathbb{V}[T]$:

$$\begin{aligned}
\mathbb{V}[T] &= \mathbb{E}[T^2] - \mathbb{E}[T]^2 \\
&= \binom{n}{3} p^3(1 + 3np^2 - 9p^2 - 3np^3 + 8p^3)
\end{aligned}$$

Note that, if $np = \omega(1)$, then $\mathbb{V}[T]/\mathbb{E}[T]^2 \to 0$ when $n \to \infty$. Chebyshev's inequality states that, for any $k > 0$,

$$\mathbb{P}[|T - \mathbb{E}[T]| \geq k\sqrt{\mathbb{V}[T]}] \leq \frac{1}{k^2}.$$

Then, for any $c > 0$, we can take $k = c\,\mathbb{E}[T]/\sqrt{\mathbb{V}[T]}$ and obtain

$$\mathbb{P}[|T - \mathbb{E}[T]| \geq c\,\mathbb{E}[T]] \leq \frac{\mathbb{V}[T]}{c^2\mathbb{E}[T]^2},$$

which vanishes for fixed c. □

We can now proceed to the proof of Theorem 4.5.

Proof of Theorem 4.5. We will first prove the result under the additional assumption that $p = \omega(\log n/n)$, which will be removed by the end of the proof.

Let $G = (V, E)$ be a $G(n, p)$ random graph, and let T be the number of triangles in G. Fix $c > 0$ arbitrary and $\delta \in (0, 1)$, and define the following sequences of events on G:

$$A_n = \{|T - \mathbb{E}[T]| < c\, \mathbb{E}[T]\},$$

$$B_n = \{\text{all degrees } \leq (n-1)p(1+\delta)\}.$$

Lemma 4.6 and Lemma 4.7 ensure that $\mathbb{P}(A_n \cap B_n) \to 1$ as $n \to \infty$. We will prove that this intersection event is contained in the event $\{G(n, p) \text{ is locally symmetric}\}$.

In A_n, there are at most $(1 + c)\binom{n}{3}p^3$ triangles in G. Each edge in a triangle appears as a peripheral edge in the closed neighborhood of its opposite vertex in this triangle. Therefore, summing over all vertices' closed neighborhoods, there are, at most, $3(1 + c)\binom{n}{3}p^3 = o(n)$ peripheral edges. This implies that, at least, $n - 3(1 + c)\binom{n}{3}p^3$ vertices have no peripheral edges in their closed neighborhoods. Let C be the set of such vertices.

In B_n, every vertex has degrees in the range $[0, (n-1)p(1 \pm \delta))$. Let D be the set of integers satisfying this property.

Now, for $p = o(n^{-2/3})$, we have:

$$\begin{aligned} |C| &= n - \theta((np)^3) \\ &= n - o(n), \\ |D| &\leq (n-1)p(1+\delta) + 1 \\ &= o(\sqrt[3]{n}). \end{aligned}$$

This implies that $|C| > |D|$ for sufficiently large n. By the pigeonhole principle, in $A_n \cap B_n$, there must be at least two vertices in C with the same degree. These vertices' closed neighborhoods are stars of the same size, therefore they must be isomorphic. This implies our result.

It still remains for us to lift the assumption that $p = \omega(\log n/n)$. The main issue to resolve is that we cannot bound directly the size of D, since Lemma 4.6 does not apply. Instead, we must try and sidestep the issue to make the argument work.

Pick some $p' \geq p$, such that $\omega(\log n/n) \leq p' \leq o(n^{-2/3})$, and consider the sequence of events:

$$B'_n = \{\text{all degrees } \leq (n-1)p'(1+\delta)\}.$$

Note that our probability measure $\mathbb{P}$ is associated with $G(n, p)$, not with $G(n, p')$. However, since B'_n happens a.a.s. under $G(n, p')$ (by Lemma 4.6), and the events

B'_n represent a decreasing property[1], it must also happen a.a.s. under $G(n,p)$, so $\mathbb{P}(B'_n) \to 1$.

Replacing B_n by B'_n in our previous argument, the conclusion again follows. □

4.3 Asymmetry regime

While the proof of this symmetry regime is rather straightforward, the identification of an asymmetry regime is much more convoluted. In this thesis, we show the following result:

Theorem 4.8. *A $G(n,p)$ random graph with $\omega(n^{-1/2+\delta_1}) \leq p \leq o(n^{-3/7-\delta_2})$ for constant $\delta_1, \delta_2 > 0$ is locally asymmetric a.a.s.*

We present here a sketch of our proof. Ultimately, we would like to show that, under the theorem's hypothesis, a $G(n,p)$ random graph is locally symmetric with probability $o(1)$.

1. By a simple union bound argument, it suffices to show that two vertices in this graph are locally symmetric with probability $o(n^{-2})$, that is, their closed neighborhoods are isomorphic with probability $o(n^{-2})$. We call these neighborhoods G_1 and G_2 for the purposes of this sketch;

2. We relate isomorphism to a distance metric over degree sequences of graphs, which leaves us to show that this distance between G_1 and G_2 is equal to 0 with probability $o(n^{-2})$;

3. Since the intersection of G_1 and G_2 is possibly non-empty, they are not necessarily independent. Removing this intersection from both graphs gives us two independent random graphs, which we call G'_1 and G'_2, respectively, and has an impact of $O(n^2p^4)$ in the distance of the degree sequences with probability higher than $1 - o(n^{-2})$. Thus, it suffices to show that the distance between G'_1 and G'_2 is $O(n^2p^4)$ with probability $o(n^{-2})$;

4. We approximate the degree sequences of G'_1 and G'_2 by two sequences D_1 and D_2 of independent random variables, with the same marginal distributions. This approximation preserves power-law decaying probabilities of events, leaving us to prove that the distance between D_1 and D_2 is $O(n^2p^4)$ with probability $o(n^{-2})$;

[1] A *decreasing property* is a property preserved under removal of edges (such as "$G(n,p)$ is not connected"). A standard coupling argument shows that, for all n, the probability of such properties is a decreasing function of p.

5. Finally, by grouping the elements of D_1 and D_2 into "buckets", we reduce the distance between these sequences to the L_1-distance of independent multinomial random vectors, which we show to be larger than $(np)^{1/2-\varepsilon}$ for any $\varepsilon > 0$, with probability $1 - o(n^{-a})$ for any $a > 0$.

To carry out this sketch, we need several additional tools, which we present in the following subsections. In subsection 4.3.1, we introduce degree sequences of graphs and the *degree sequence edit distance*, and present some basic properties. In subsection 4.3.2, we present a framework for approximating the degree sequence of a $G(n, p)$ random graph by a sequence of independent random variables, and extend this framework to handle two random graphs simultaneously. This framework will be used to characterize the degree sequence edit distance between two random graphs in subsection 4.3.3. Finally, this result and some additional ones will be applied in subsection 4.3.4, where we convert our sketch into a proof for Theorem 4.8.

4.3.1 Degree sequences

The first and most important concept is that of the degree sequence of a graph, which we apply in step 2 of our sketch. We present this concept in a slightly different form:

Definition 4.9. *For any graph $G = (V, E)$, the* degree function *of G is the function $\phi_G : \mathbb{N}_0 \to \mathbb{N}_0$ such that $\phi_G(k) = |\{v \in V : d_G(v) = k\}|$ for all $k \in \mathbb{N}_0$.*

The degree function of G simply returns, for an input k, the number of vertices with degree k in G. This makes it equivalent to the degree sequence of G, whenever the listing order of the degrees is irrelevant or deterministically given.

To identify asymmetry regimes, we must identify conditions under which no two vertices in a $G(n, p)$ random graph are symmetric a.a.s. The asymmetry of vertices is defined as the lack of an isomorphism between their closed neighborhoods, which relates to degree functions via the following remark:

Remark 4.10. *For G and G' isomorphic graphs, $\phi_G \equiv \phi_{G'}$.*

To enable a more fine-grained look into these closed neighborhoods, we need one additional definition:

Definition 4.11. *Let $G = (V, E)$ and $G' = (V', E')$ be two graphs. The* degree sequence edit distance *between G and G' (denoted by $\Delta(G, G')$) is given by:*

$$\Delta(G, G') = \sum_k |\phi_G(k) - \phi_{G'}(k)|.$$

Remark 4.12. *Let μ be the counting measure on $\mathbb{N}$. Then:*

$$\Delta(G, G') = \int |\phi_G - \phi_H| \, \mathrm{d}\mu = \|\phi_G - \phi_H\|_1.$$

Thus Δ is a semimetric over the space of all graphs, with $\Delta(G, G') = 0$ iff $\phi_G \equiv \phi_{G'}$.

Remark 4.13. *For G and G' isomorphic graphs, $\Delta(G, G') = 0$.*

The degree sequence edit distance can be thought of as being an edit distance by operations over the strings generated by the degree sequences of G and G', where operations of insertion and deletion are counted to the distance, but reorderings are not and can be used freely. In this matter, it differs fundamentally from traditional distance measures such as the Hamming and Levenshtein distances [34], in which the order of elements in the string is fundamental. It can also be interpreted as the size of the symmetric difference of the multisets generated by the degree sequences of graphs G and G'.

To formalize these interpretations, consider any partial mapping between the vertex sets of two graphs. We can measure the *degree mismatch count* of this mapping, which is a simple count of vertices, from both graphs, that are either mapped to vertices with different degrees or left unmapped.

Definition 4.14. *Let $G = (V, E)$, $G' = (V', E')$ be two graphs, and let $f : S \to S'$ be a bijective function from $S \subseteq V$ to $S' \subseteq V'$. The* degree mismatch count *of f (denoted by δ_f) is given by*

$$\begin{aligned}\delta_f = &|\{v \in S \ : \ d_G(v) \neq d_{G'}(f(v))\}| + |V \setminus S| \\ &+ |\{v' \in S' \ : \ d_{G'}(v') \neq d_G(f^{-1}(v'))\}| + |V' \setminus S'|.\end{aligned}$$

The degree sequence edit distance between two graphs is, then, the smallest possible degree mismatch count between their vertex sets:

Theorem 4.15. *Let $G = (V, E)$ and $G' = (V', E')$ be two graphs. Then:*

$$\Delta(G, G') = \min_{\substack{f \ : \ S \ \to \ S' \\ f \text{ bijective}, S \subseteq V, S' \subseteq V'}} \delta_f.$$

Proof. The statement follows from inequalities on both directions. We begin by showing $\Delta(G, G') \geq \min_f \delta_f,$. It is enough to construct a function g such that $\Delta(G, G') = \delta_g$, and we perform this construction by "slices", one for each possible vertex degree.

For each $k \in \mathbb{N}_0$, let $v_1^k, \ldots, v_{\phi_G(k)}^k$ and $w_1^k, \ldots, w_{\phi_{G'}(k)}^k$ be enumerations of degree-k vertices in G and G', respectively. Write $m_k = \min(\phi_G(k), \phi_{G'}(k))$,

$V_k = \{v_1^k, \ldots, v_{m_k}^k\}$ and $V'_k = \{w_1^k, \ldots, w_{m_k}^k\}$, and construct function $g_k : V_k \to V'_k$ mapping v_j^k to w_j^k, for $j = 1, \ldots, m_k$. Note that g_k leaves $|\phi_G(k) - \phi_{G'}(k)|$ degree-k nodes unmapped, all from V (if $\phi_G(k) \geq \phi_{G'}(k)$) or from V' (if $\phi_G(k) \leq \phi_{G'}(k)$), and which, by construction, cannot be mapped by any other function $g_{k'}$.

Now, construct function $g : \cup_k V_k \to \cup_k V'_k$ as $g = \cup_k g_k$. By double counting the number of nodes left unmapped by g (from both V and V'), we see that this number is equal to $|V \setminus (\cup_k V_k)| + |V' \setminus (\cup_k V'_k)|$ by definition, and to $\sum_k |\phi_G(k) - \phi_{G'}(k)|$ by construction. Furthermore, our construction also ensures that for every node mapped by g with degree k in G, its image has degree k in G, and vice-versa for nodes in G'. Therefore, it holds that $|\{v \in \cup_k V_k \ : \ d_G(v) \neq d_{G'}(g(v))\}| = |\{v \in \cup_k V'_k \ : \ d_{G'}(v) \neq d_G(g^{-1}(v))\}| = 0$, and:

$$
\begin{aligned}
\Delta(G, G') &= \sum_k |\phi_G(k) - \phi_{G'}(k)| \\
&= \sum_k |\phi_G(k) - \phi_{G'}(k)| + 0 \\
&= |V \setminus (\cup_k V_k)| + |V' \setminus (\cup_k V'_k)| \\
&\quad + |\{v \in \cup_k V_k \ : \ d_G(v) \neq d_{G'}(g(v))\}| \\
&\quad + |\{v' \in \cup_k V'_k \ : \ d_{G'}(v') \neq d_G(g^{-1}(v'))\}| \\
&= \delta_g.
\end{aligned}
$$

Now, it only remains to show $\Delta(G, G') \leq \min_f \delta_f$. We will show that $\Delta(G, G') \leq \delta_h$ for any partial mapping h, and we will again proceed by "slices" in our argument. Let $h : S \to S'$ be an arbitrary bijection with $S \subseteq V$ and $S' \subseteq V'$.

Consider all vertices with degree k in both G and G'. If $\phi_G(k) \geq \phi_{G'}(k)$, then at most $\phi_{G'}(k)$ k-degree vertices in G can be mapped by h into k-degree nodes in G'. This implies that at least $\phi_G(k) - \phi_{G'}(k)$ k-degree vertices in G must either be left unmapped by g (thus belonging to $V \setminus S$) or be mapped to vertices in G' with degree different than k (and therefore belonging to $\{v \in \cup_k V_k \ : \ d_G(v) \neq d_{G'}(h(v))\}$). Analogously, if $\phi_G(k) \leq \phi_{G'}(k)$, at least $\phi_{G'}(k) - \phi_G(k)$ k-degree vertices in G' must either be left unmapped by g (this time, belonging in $V' \setminus S'$) or be mapped to vertices in G with degree different than k.

In both cases, there is a contribution of $|\phi_G(k) - \phi_{G'}(k)|$ to δ_h coming exclusively from nodes of degree k in G and G'. Putting together the disjoint contributions from each "slice" ensures that $\delta_h \geq \sum_k |\phi_G(k) - \phi_{G'}(k)| = \Delta(G, G')$. Since h was arbitrary, $\min_h \delta_h \geq \Delta(G, G')$, as desired. □

This property can be used to relate the degree functions of a graph and its subgraphs. For any set of vertices $S \subseteq V$, denote by $G[S]$ the subgraph of G

induced by S, and by $C(S)$ the set of edges with one endpoint in S and another in $V \setminus S$.

Corollary 4.16. *For any graph $G = (V, E)$ and any $S \subset V$,*

$$\Delta(G, G[S]) \leq |V \setminus S| + |C(S)|.$$

Proof. By virtue of Theorem 4.15, it is enough to show that $\delta_g = |V \setminus S| + |C(S)|$ for some partial mapping g between G and $G[S]$.

Take $g : S \to S$ to be the identity function on S. Then g is a partial mapping between the vertex sets of G and $G[S]$, and its degree mismatch count is given by:

$$\delta_g = |\{v \in G[S] \ : \ d_{G[S]}(v) \neq d_G(v)\}| + |V \setminus S|.$$

Now, notice that, for any vertex $v \in S$, $d_{G[S]}(v) \neq d_G(v)$ if and only if v is adjacent to some vertex outside S. Since there are $|C(S)|$ edges between S and $V \setminus S$, there can be at most $|C(S)|$ such vertices, and the result follows. □

This result will be used in step 3 of our proof.

4.3.2 Approximation framework

Analysis of the degree sequences of $G(n, p)$ random graphs takes place in steps 4 and 5 of our proof sketch, in which we, respectively, show a technique for handling these degree sequences and apply it to the degree sequence edit distance, presented in subsection 4.3.1. The main result in this section is the following approximation theorem:

Theorem 4.17. *Let $\vec{D}_n^{(1)}$ and $\vec{D}_n^{(2)}$ be the degree sequences of two independent $G(n, p)$ random graphs, with probability distribution $\mathbb{P}_{\mathcal{D}_n}$.*

Furthermore, let $\mathcal{F}_n$ be the σ-algebra generated by $\vec{D}_n^{(1)}$ and $\vec{D}_n^{(2)}$, and let $\mathbb{P}_{\mathcal{B}_n}$ be a probability measure under which $\vec{D}_n^{(1)}$ and $\vec{D}_n^{(2)}$ are random vectors with n independent coordinates, each having distribution $\mathrm{Bin}(n-1, p)$.

Then, for any sequence of events A_n measurable under $\mathcal{F}_n^{\otimes 2}$ and any fixed $a > 0$, if $\mathbb{P}_{\mathcal{B}_n}^{\otimes 2}(A_n) = o(n^{-a})$, then $\mathbb{P}_{\mathcal{D}_n}^{\otimes 2}(A_n) = o(n^{-a})$.

In a nutshell, Theorem 4.17 allows us to consider the degree sequence of two $G(n, p)$ random graphs as sequences of independent random variables, without interfering with power-law decays in the probability of events on this model. The proof of this result is rather long and technical, but paramount to the execution of our main proof.

Obtaining a precise characterization of the degree sequence of a $G(n, p)$ random graph has been one of the toughest challenges in understanding its structure. The

main reason for this is that, even though the degrees of any two specific nodes are only mildly correlated (due to the possible edge between them), it is still a nontrivial task to compose these correlations into a manageable joint distribution for the degrees. Most results on this matter address the distribution of the t-th largest degree, for some $t(n)$ generally bounded.

More recently, though, a framework has been set by McKay and Wormald [35] for approximating the degree sequence by a sequence of independent random variables, with tight bounds on the error of the probabilities of events estimated by this approximation. This framework has been successfully applied in several contexts. For instance, Kostochka and West [36] use it to analyze the middle degree asymptotics of random graphs, which relates to Chvátal's condition for Hamiltonian graphs, and Skerman [37] applies a similar technique to analyze degrees in a random bipartite graph model. Let us detail this framework a bit further before proceeding.

It is quite intuitive to say that the degree sequence of a $G(n,p)$ random graph is similar to a sequence of independent random variables, each having distribution $\mathrm{Bin}(n-1,p)$. Their framework formalize this via several theorems and lemmas, each performing one of four steps in the approximation process that we will detail. Notation will be kept as similar as possible to the original work [35].

For some fixed $n \in \mathbb{N}$, take the set $I_n = \{0,\dots,n-1\}^n$ equipped with the discrete σ-algebra as our measurable space. Let $d = (d_1,\dots,d_n)$ be some element in this space. Also, let $p = p(n) \in (0,1)$, and denote $N = \binom{n}{2}$ and $q = 1-p$.

In the *binomial model* $\mathcal{B}_{n,p}$, d is distributed as a sequence of n independent $\mathrm{Bin}(n-1,p)$ random variables. This can be achieved by evaluating d under the probability measure $\mathbb{P}_{\mathcal{B}_{n,p}} = \mathrm{Bin}(n-1,p)^{\otimes n}$.[2] We would like to assert that this model is similar to the degree sequence of a $G(n,p)$ random graph. We call this the *degree sequence model* ($\mathcal{D}_{n,p}$), and denote by $\mathbb{P}_{\mathcal{D}_{n,p}}$ the probability measure under which d has this distribution. Note that the sum of degrees in any graph is necessarily even, which means d will take, with probability 1, values on the set $E_n = \{d \in I_n \,:\, M(d) \text{ is even}\}$ (where $M = M(d) = \|d\|_1$ is the sum of the components of d).

The approximation process requires three additional models (with corresponding probability measures) that will perform a transition from the binomial model to the degree sequence model, with two of them making d acquire properties from the degree sequence model that are not present in the binomial model, and the third one acting as a technical middleman. The first model is the *even-sum binomial*

[2]Given a probability measure $\mathcal{P}$, we denote by $\mathcal{P}\otimes\mathcal{P}$ the product probability measure (on the product space) given by taking the product of two copies of $\mathcal{P}$. A direct interpretation of this space is that, if X is a random element with distribution $\mathcal{P}$, then a random element with distribution $\mathcal{P}\otimes\mathcal{P}$ is an 2-dimensional vector with mutually independent coordinates and each of them distributed as X. For the product of n copies of $\mathcal{P}$, we write $\mathcal{P}^{\otimes n} = \mathcal{P}\otimes\mathcal{P}\otimes\ldots\otimes\mathcal{P}$. In this case, all coordinates are mutually independent and have distribution $\mathcal{P}$.

model ($\mathcal{E}_{n,p}$). It ensures that d indeed takes values in E_n with probability one. To ensure minimum distortion between probability of elements of E_n, this model is simply set to be the restriction of the binomial model to the set E_n.[3] Then, the *weighted even-sum binomial model* ($\mathcal{E}'_{n,p}$) ensures the stronger property that M has the same distribution as it does under the degree sequence model (namely, that $M/2$ is distributed as $\mathrm{Bin}(N,p)$). To insert as little interference as possible into the relative probabilities of any two points in E_n, the probabilities of all points E_n are rescaled (or *reweighted*) uniformly on each set $S_m = \{d \in E_n \ : \ M(d) = m\}$,to make these sets have the desired probability.

To perform the bridge between $\mathcal{E}'_{n,p}$ and $\mathcal{E}_{n,p}$, they have introduced the *integrated model* $\mathcal{I}_{n,p}$, which is essentially a "noisy" version of the even-sum model $\mathcal{E}_{n,p}$. The model $\mathcal{I}_{n,p}$ is obtained from $\mathcal{E}_{n,p}$ by switching from a fixed parameter p to a random parameter p' that quickly concentrates around p. More specifically, p' must be distributed as a truncated normal variable, with expected value p, variance $pq/2N$, and restricted to the unit interval.

We can informally summarize the approximation scheme as follows:

$$\mathbb{P}_{\mathcal{B}_{n,p}} \approx \mathbb{P}_{\mathcal{E}_{n,p}} \approx \mathbb{P}_{\mathcal{I}_{n,p}} \approx \mathbb{P}_{\mathcal{E}'_{n,p}} \approx \mathbb{P}_{\mathcal{D}_{n,p}}$$

Now, for these approximations to work, it is necessary for $p(n)$ to lie in a "good behavior range", in which case $p = p(n)$ is said to be *acceptable*. The last approximation, in particular, is hard to tighten in general, so the necessary conditions for this approximation to work are brought into the definition of acceptable function:

Definition 4.18. *Let $\lambda = \lambda(d) = M(d)/2N$ and $\gamma_2 = \gamma_2(d) = (n-1)^{-2}\sum_{i=1}^{n}(d_i - M(d))^2$. A function $p = p(n)$ is acceptable if the following conditions hold:*

1. *$pqN = \omega(n)\log n$;*
2. *there are sets $R_p(n) \subset E_n$ and a real function $\delta(n) = o(1)$ such that:*
   (a) *$\mathbb{P}_{\mathcal{D}_{n,p}}(R_p(n)), \mathbb{P}_{\mathcal{E}_{n,p}}(R_p(n)) = 1 - n^{-\omega(n)}$;*
   (b) *for every $d \in R_p(n)$, there is some δ_d such that $|\delta_d| \leq \delta(n)$ and*

$$\frac{\mathbb{P}_{\mathcal{D}_{n,p}}(d)}{\mathbb{P}_{\mathcal{E}'_{n,p}}(d)} = \exp\left\{\frac{1}{4}\left(1 - \frac{\gamma_2^2}{\lambda^2(1-\lambda)^2}\right)\right\} \cdot \exp\{\delta_d\}.$$

The second condition in this definition requires a set $R_p(n)$ to exist in our sample space E_n, with very large probability in $\mathcal{D}_{n,p}$ and $\mathcal{E}_{n,p}$ (the probability of its complement in both models vanishes faster than any standard exponential), in which the

[3]That is, the corresponding probability measure is the measure for the binomial model conditional to the event E_n, evaluated only on the events E_n.

models $\mathcal{D}_{n,p}$ and $\mathcal{E}'_{n,p}$ uniformly agree to a ratio that approaches 1. This condition is required for the proofs to be carried out, though it has been conjectured by McKay and Wormald that condition 1 in the definition is sufficient for $p(n)$ to be acceptable — to the best of our knowledge, this conjecture is still open. For our purposes, they have identified an interesting regime for $p(n)$ in which these conditions hold [35]:

Theorem 4.19. *$p(n)$ is acceptable whenever $\omega(n)\log n/n^2 \leq pq \leq o(n^{-1/2})$.*

The execution of this approximation scheme has been broken down into a number of pieces with various levels of complexity, so to fit different possibilities of applications. In our particular case, we would like to ensure that this scheme is well-suited for approximating probabilities that vanish faster than power laws in n. For this purpose, we extract the following results from [35], condensed in a single theorem.

Theorem 4.20. *Write $\phi(x;\mu,\sigma^2)$ the density function of the normal distribution, and $V_{n,p} = \int_0^1 \phi(x;p,pq/2N)\mathrm{d}x$. Then the following statements hold:*

1. *For any event $A_n \subseteq E_n$,*

$$\mathbb{P}_{\mathcal{E}_{n,p}}(A_n) = \frac{2\mathbb{P}_{\mathcal{B}_{n,p}}(A_n)}{1+(q-p)^{2N}};$$

2. *For any event $A_n \subseteq E_n$,*

$$\mathbb{P}_{\mathcal{I}_{n,p}}(A_n) = \frac{1}{V_{n,p}}\int_0^1 \phi(x;p,pq/2N)\mathbb{P}_{\mathcal{E}_{n,x}}(A_n)\mathrm{d}x;$$

3. *If $pqN \to \infty$ and $y = y(n) = o(\sqrt[6]{pqN})$, then*

$$\mathbb{P}_{\mathcal{I}_{n,p}}(d) = \mathbb{P}_{\mathcal{E}'_{n,p}}(d)\left(1 + O\left(\frac{1+|y|^3}{\sqrt{pqN}}\right)\right)$$

   uniformly over $\{d \in E_n \ : \ |M(d)-2Np| \leq 2y\sqrt{Npq}\}$;

4. *If $\omega(n)\log n/n^2 \leq pq \leq o(n^{-1/2})$, then there are sets $R_p(n), R'_p(n) \subseteq E_n$ and a real function $\delta(n) = o(1)$ such that:*

   (a) *$\mathbb{P}_{\mathcal{D}_{n,p}}(R_p(n)), \mathbb{P}_{\mathcal{D}_{n,p}}(R'_p(n)) = 1 - n^{-\omega(n)}$;*

   (b) *in $R'_p(n)$, $\gamma_2 = \lambda(1-\lambda)(1+o(1))$;*

   (c) *for every $d \in R_p(n)$, there is some δ_d such that $|\delta_d| \leq \delta(n)$ and*

$$\frac{\mathbb{P}_{\mathcal{D}_{n,p}}(d)}{\mathbb{P}_{\mathcal{E}'_{n,p}}(d)} = \exp\left\{\frac{1}{4}\left(1 - \frac{\gamma_2^2}{\lambda^2(1-\lambda)^2}\right)\right\}\cdot\exp\{\delta_d\}.$$

Proof. All referenced results used in this proof have been extracted from [35], to which we refer the reader for notation and statements. Statement 1 is a particular case of corollary 4.3 taking $f = \mathbb{1}_{A_n}$ the indicator function of the event A_n, simplified by theorem 4.2 and the observation that, since $f = 0$ in $I_n \setminus E_n$, $f = \tilde{f}$. Statement 2 is a rewriting of lemma 2.4, consequence of the construction of $\mathbb{P}_{\mathcal{I}_{n,p}}$ from $\mathbb{P}_{\mathcal{E}_{n,p}}$ and an application of the law of total probability — we note that, for $x \in [0,1]$, $\phi(x; p, pq/2N)/V_{n,p}$ is the density function of the random parameter p' used in the construction. Statement 3 simply restates theorem 3.6. Statement 4 comes from the definition of acceptability and corollary 3.5, noting that the hypothesis implies $p(n)$ is acceptable. □

Theorem 4.20 suffices for us to ensure suitability of the scheme to our purposes, as long as we restrict ourselves to properties (i.e. events) on a single $G(n,p)$ random graph:

Theorem 4.21. *Let A_n be a sequence of events in E_n. If p satisfies $\omega(\log n/n) \leq p \leq o(n^{-1/2})$, then for any fixed $a > 0$, $\mathbb{P}_{\mathcal{B}_{n,p}}(A_n) = o(n^{-a})$ implies $\mathbb{P}_{\mathcal{D}_{n,p}}(A_n) = o(n^{-a})$.*

Each step in the proof of this theorem is a simplified version of the corresponding step in the proof of Theorem 4.23, which considers two independent random graphs and will be presented later. For brevity, we will explicitly provide proof only for the latter theorem.

It is worth noting that, in the abstract of their paper, McKay and Wormald state that their techniques can be used to determine highly accurate asymptotics for probabilities that, in $\mathcal{D}_{n,p}$, are greater than any (fixed) power law. While this statement, and particularly the meaning of "highly accurate", has been posed in an informal fashion, it can be understood, in light of their approximation scheme, to mean the following formal statement:

> If $p(n)$ is acceptable, then for any event $A_n \subseteq E_n$, if $\mathbb{P}_{\mathcal{D}_{n,p}}(A_n) = \omega(n^{-k})$ for some fixed k, then there are functions $\alpha(n) = \theta(1)$ and $\varepsilon(n) = n^{-\omega(n)}$ such that $\mathbb{P}_{\mathcal{D}_{n,p}}(A_n) = \mathbb{P}_{\mathcal{B}_{n,p}}(A_n) \cdot \alpha(n) + \varepsilon(n)$.

Theorem 4.21 could, in principle, be derived as a corollary of this fact since it asserts that, if $p(n)$ is acceptable, $\mathbb{P}_{\mathcal{D}_{n,p}}(A_n) = \Omega(n^{-k})$ implies $\mathbb{P}_{\mathcal{B}_{n,p}}(A_n) = \Omega(n^{-k})$, which is the contrapositive of our result. However, since this fact has neither been formally stated nor proven, we choose to carry out a proof directly, by using the same fundamental techniques (i.e., individual pieces of the approximation scheme) that would be required to prove the aforementioned fact.

In comparing the degree sequences of two independent random graphs, this framework cannot be applied directly, though it would seem trivial that, since the

two degree sequences are independent of each other and can both be individually approximated by i.i.d. sequences with small errors on the corresponding probabilities of events, the joint approximation of both degree sequences should yield a small error as well. However, we find it essential that this extension of the single-graph case be obtained formally. As we see in what follows, even though such extension is indeed possible, achieving it is far from trivial, since we must deal with a product space containing events which compare the structure of both graphs.

Before we proceed, let us introduce some notation. For $p, p' \in (0,1)$, denote by $\mathbb{P}_{\mathcal{B}_{n,p,p'}}$ the probability measure $\mathbb{P}_{\mathcal{B}_{n,p}} \otimes \mathbb{P}_{\mathcal{B}_{n,p'}}$ over I_n^2 — and similarly for measures in other models, over E_n^2. Our goal is to perform the following approximation scheme:

$$\mathbb{P}_{\mathcal{B}_{n,p,p'}} \approx \mathbb{P}_{\mathcal{E}_{n,p,p'}} \approx \mathbb{P}_{\mathcal{I}_{n,p,p'}} \approx \mathbb{P}_{\mathcal{E}'_{n,p,p'}} \approx \mathbb{P}_{\mathcal{D}_{n,p,p'}}$$

Let us stress that $\mathbb{P}_{\mathcal{D}_{n,p,p'}} = \mathbb{P}_{\mathcal{D}_{n,p}} \otimes \mathbb{P}_{\mathcal{D}_{n,p'}}$ is the distribution of the degree sequence of two random graphs $G(n,p)$ and $G(n,p')$ independent of each other, and $\mathbb{P}_{\mathcal{B}_{n,p,p'}}$ is the corresponding approximation by two independent sequences of i.i.d. random variables.

We will extend our notation further and write $q' = 1 - p'$, and denote by $d \times d'$ some element of I_n^2. We will also write $\lambda = \lambda(d \times d') = M(d)/2N$, $\lambda' = \lambda'(d \times d') = M(d')/2N$, $\gamma_2 = \gamma_2(d \times d') = (n-1)^{-2} \sum_{i=1}^n (d_i - M(d))^2$ and $\gamma_2' = \gamma_2'(d \times d') = (n-1)^{-2} \sum_{i=1}^n (d_i' - M(d'))^2$. For the models we have presented, the following extension of Theorem 4.20 holds:

Theorem 4.22. *Let $\phi(x; \mu, \sigma^2)$ and $V_{n,p}$ as in Theorem 4.20. Then the following statements hold:*

1. *For any event $A_n \subseteq E_n^2$,*

$$\mathbb{P}_{\mathcal{E}_{n,p,p'}}(A_n) = \frac{4\mathbb{P}_{\mathcal{B}_{n,p,p'}}(A_n)}{[1 + (q-p)^{2N}][1 + (q'-p')^{2N}]};$$

2. *For any event $A_n \subseteq E_n^2$,*

$$\mathbb{P}_{\mathcal{I}_{n,p,p'}}(A_n) = \frac{1}{V_{n,p} V_{n,p'}} \int_0^1 \int_0^1 \phi(x; p, \frac{pq}{2N}) \phi(x'; p', \frac{p'q'}{2N}) \mathbb{P}_{\mathcal{E}_{n,x,x'}}(A_n) \mathrm{d}x \mathrm{d}x';$$

3. *If $pqN, p'q'N \to \infty$ and $y = y(n)$ is both $o(\sqrt[6]{pqN})$ and $o(\sqrt[6]{p'q'N})$, then*

$$\mathbb{P}_{\mathcal{I}_{n,p,p'}}(d \times d') = \mathbb{P}_{\mathcal{E}'_{n,p,p'}}(d \times d') \left(1 + O\left(\frac{1 + |y|^3}{\sqrt{pqN}}\right) + O\left(\frac{1 + |y|^3}{\sqrt{p'q'N}}\right)\right)$$

uniformly over $\{d \times d' \in E_n^2 \ : \ |M(d) - 2Np| \leq 2y\sqrt{Npq}, |M(d') - 2Np'| \leq 2y\sqrt{Np'q'}\}$;

4. *If* $\omega(n) \log n/n^2 \leq pq, p'q' \leq o(n^{-1/2})$, *then there are sets* $S_{p,p'}(n) \subseteq E_n^2$ *and* $S'_{p,p'}(n) \subseteq E_n^2$ *and a real function* $\varepsilon(n) = o(1)$ *such that:*

   (a) $\mathbb{P}_{\mathcal{D}_{n,p,p'}}(S_{p,p'}(n)), \mathbb{P}_{\mathcal{D}_{n,p,p'}}(S'_{p,p'}(n)) = 1 - n^{-\omega(n)}$;

   (b) *in* $S'_{p,p'}(n)$, $\gamma_2 = \lambda(1-\lambda)(1+o(1))$ *and* $\gamma'_2 = \lambda'(1-\lambda')(1+o(1))$;

   (c) *for every* $d \times d' \in S_{p,p'}(n)$, *there is some* $\varepsilon_{d\times d'}$ *such that* $|\varepsilon_{d\times d'}| \leq \varepsilon(n)$ *and*

$$\frac{\mathbb{P}_{\mathcal{D}_{n,p,p'}}(d \times d')}{\mathbb{P}_{\mathcal{E}'_{n,p,p'}}(d \times d')} = \exp\left\{\frac{1}{4}\left(2 - \frac{\gamma_2^2}{\lambda^2(1-\lambda)^2} - \frac{(\gamma'_2)^2}{(\lambda')^2(1-\lambda')^2}\right)\right\} \cdot \exp\{\varepsilon_{d\times d'}\}.$$

Proof. **Statement 1** Let $\mathcal{F}$ be the family of subsets of E_n^2 for which the statement's equality holds. We will prove that (i) $\mathcal{F}$ contains all rectangles (i.e., events of the form $B \times B'$, with B, $B' \subset E_n$) and (ii) $\mathcal{F}$ is a λ-system. This is enough since, by Dynkin's theorem, $\mathcal{F}$ must contain the σ-algebra generated by the rectangles, which is the discrete σ-algebra over E_n^2.

For the first claim, for any rectangle $A = B \times B'$ in E_n^2, by Theorem 4.20(1), we have that

$$\begin{aligned}
\mathbb{P}_{\mathcal{E}_{n,p,p'}}(A) &= \mathbb{P}_{\mathcal{E}_{n,p}} \otimes \mathbb{P}_{\mathcal{E}_{n,p'}}(B \times B') \\
&= \mathbb{P}_{\mathcal{E}_{n,p}}(B)\mathbb{P}_{\mathcal{E}_{n,p'}}(B') \\
&= \frac{2\mathbb{P}_{\mathcal{B}_{n,p}}(B)}{1 + (q-p)^{2N}} \cdot \frac{2\mathbb{P}_{\mathcal{B}_{n,p'}}(B')}{1 + (q'-p')^{2N}} \\
&= \frac{4\mathbb{P}_{\mathcal{B}_{n,p}}(B)\mathbb{P}_{\mathcal{B}_{n,p'}}(B')}{[1 + (q-p)^{2N}][1 + (q'-p')^{2N}]} \\
&= \frac{4\mathbb{P}_{\mathcal{B}_{n,p,p'}}(A)}{[1 + (q-p)^{2N}][1 + (q'-p')^{2N}]}.
\end{aligned}$$

Therefore, $\mathcal{F}$ contains all rectangles.

For the second claim, note that $\mathcal{F}$ contains E_n^2, since it is a rectangle; $\mathcal{F}$ is

closed by complements, since for any $A \in \mathcal{F}$, it holds that

$$\begin{aligned}
&\mathbb{P}_{\mathcal{E}_{n,p,p'}}(E_n^2 \setminus A) \\
&= \mathbb{P}_{\mathcal{E}_{n,p,p'}}(E_n^2) - \mathbb{P}_{\mathcal{E}_{n,p,p'}}(A) \\
&= \frac{4\mathbb{P}_{\mathcal{B}_{n,p,p'}}(E_n^2)}{[1+(q-p)^{2N}][1+(q'-p')^{2N}]} - \frac{4\mathbb{P}_{\mathcal{B}_{n,p,p'}}(A)}{[1+(q-p)^{2N}][1+(q'-p')^{2N}]} \\
&= \frac{4\mathbb{P}_{\mathcal{B}_{n,p,p'}}(E_n^2 \setminus A)}{[1+(q-p)^{2N}][1+(q'-p')^{2N}]}
\end{aligned}$$

and $E_n^2 \setminus A \in \mathcal{F}$; and $\mathcal{F}$ is also closed by disjoint enumerable unions, since for any sequence $B_1, B_2, \ldots$ in $\mathcal{F}$, if $B_1, B_2, \ldots$ are disjoint, then

$$\begin{aligned}
\mathbb{P}_{\mathcal{E}_{n,p,p'}}\left(\biguplus_{i=1}^{\infty} B_i\right) &= \sum_{i=1}^{\infty} \mathbb{P}_{\mathcal{E}_{n,p,p'}}(B_i) \\
&= \sum_{i=1}^{\infty} \frac{4\mathbb{P}_{\mathcal{B}_{n,p,p'}}(B_i)}{[1+(q-p)^{2N}][1+(q'-p')^{2N}]} \\
&= \frac{4\sum_{i=1}^{\infty} \mathbb{P}_{\mathcal{B}_{n,p,p'}}(B_i)}{[1+(q-p)^{2N}][1+(q'-p')^{2N}]} \\
&= \frac{4\mathbb{P}_{\mathcal{B}_{n,p,p'}}(\uplus_{i=1}^{\infty} B_i)}{[1+(q-p)^{2N}][1+(q'-p')^{2N}]}
\end{aligned}$$

and $\uplus_{i=1}^{\infty} B_i \in \mathcal{F}$. Since $\mathcal{F}$ fits the three requirements, by definition, $\mathcal{F}$ is a λ-system.

Statement 2 We will use the same strategy as in statement 1. Let $\mathcal{G}$ be the family of subsets of E_n^2 for which the statement's equality is true. First, take an arbitrary rectangle $A = B \times B'$ in E_n^2. Using Theorem 4.20(2) yields

$$\begin{aligned}
\mathbb{P}_{\mathcal{I}_{n,p,p'}}(A) &= \mathbb{P}_{\mathcal{I}_{n,p}} \otimes \mathbb{P}_{\mathcal{I}_{n,p'}}(B \times B') \\
&= \mathbb{P}_{\mathcal{I}_{n,p}}(B)\mathbb{P}_{\mathcal{I}_{n,p'}}(B') \\
&= \frac{1}{V_{n,p}} \int_0^1 \phi(x; p, pq/2N)\mathbb{P}_{\mathcal{E}_{n,x}}(B)\mathrm{d}x \\
&\quad \cdot \frac{1}{V_{n,p'}} \int_0^1 \phi(x; p', p'q'/2N)\mathbb{P}_{\mathcal{E}_{n,x'}}(B')\mathrm{d}x' \\
&= \frac{1}{V_{n,p}V_{n,p'}} \int_0^1 \int_0^1 \phi(x; p, \frac{pq}{2N})\phi(x'; p', \frac{p'q'}{2N})\mathbb{P}_{\mathcal{E}_{n,p}}(B)\mathbb{P}_{\mathcal{E}_{n,p'}}(B')\mathrm{d}x\mathrm{d}x' \\
&= \frac{1}{V_{n,p}V_{n,p'}} \int_0^1 \int_0^1 \phi(x; p, \frac{pq}{2N})\phi(x'; p', \frac{p'q'}{2N})\mathbb{P}_{\mathcal{E}_{n,x,x'}}(B)\mathrm{d}x\mathrm{d}x',
\end{aligned}$$

which means $\mathcal{G}$ contains all rectangles.

Secondly, $\mathcal{G}$ satisfies the three requirements of the definition of λ-systems: it contains E_n^2, since it is a rectangle; it is closed under complements, since for any $A \in \mathcal{G}$,

$$\begin{aligned}
\mathbb{P}_{\mathcal{I}_{n,p,p'}}(E_n^2 \setminus A) &= \mathbb{P}_{\mathcal{I}_{n,p,p'}}(E_n^2) - \mathbb{P}_{\mathcal{I}_{n,p,p'}}(A) \\
&= \frac{1}{V_{n,p}V_{n,p'}} \int_0^1 \int_0^1 \phi(x; p, \frac{pq}{2N})\phi(x'; p', \frac{p'q'}{2N}) \mathbb{P}_{\mathcal{E}_{n,x,x'}}(E_n^2) \mathrm{d}x\mathrm{d}x' \\
&\quad - \frac{1}{V_{n,p}V_{n,p'}} \int_0^1 \int_0^1 \phi(x; p, \frac{pq}{2N})\phi(x'; p', \frac{p'q'}{2N}) \mathbb{P}_{\mathcal{E}_{n,x,x'}}(A) \mathrm{d}x\mathrm{d}x' \\
&= \frac{1}{V_{n,p}V_{n,p'}} \int_0^1 \int_0^1 \Big[\phi(x; p, \frac{pq}{2N})\phi(x'; p', \frac{p'q'}{2N}) \mathbb{P}_{\mathcal{E}_{n,x,x'}}(E_n^2) \\
&\qquad - \phi(x; p, \frac{pq}{2N})\phi(x'; p', \frac{p'q'}{2N}) \mathbb{P}_{\mathcal{E}_{n,x,x'}}(A) \Big] \mathrm{d}x\mathrm{d}x' \\
&= \frac{1}{V_{n,p}V_{n,p'}} \int_0^1 \int_0^1 \phi(x; p, \frac{pq}{2N})\phi(x'; p', \frac{p'q'}{2N}) \\
&\qquad \cdot [\mathbb{P}_{\mathcal{E}_{n,x,x'}}(E_n^2) - \mathbb{P}_{\mathcal{E}_{n,x,x'}}(A)] \mathrm{d}x\mathrm{d}x' \\
&= \frac{1}{V_{n,p}V_{n,p'}} \int_0^1 \int_0^1 \phi(x; p, \frac{pq}{2N})\phi(x'; p', \frac{p'q'}{2N}) \mathbb{P}_{\mathcal{E}_{n,x,x'}}(E_n^2 \setminus A) \mathrm{d}x\mathrm{d}x'
\end{aligned}$$

and $E_n^2 \setminus A \in \mathcal{G}$; and $\mathcal{F}$ is also closed by disjoint enumerable unions, since for any sequence $B_1, B_2, \ldots$ in $\mathcal{G}$, if $B_1, B_2, \ldots$ are disjoint, then

$$\begin{aligned}
&\mathbb{P}_{\mathcal{I}_{n,p,p'}} \left(\biguplus_{i=1}^{\infty} B_i \right) \\
&= \sum_{i=1}^{\infty} \mathbb{P}_{\mathcal{I}_{n,p,p'}}(B_i) \\
&= \sum_{i=1}^{\infty} \frac{1}{V_{n,p}V_{n,p'}} \int_0^1 \int_0^1 \phi(x; p, \frac{pq}{2N})\phi(x'; p', \frac{p'q'}{2N}) \mathbb{P}_{\mathcal{E}_{n,x,x'}}(B_i) \mathrm{d}x\mathrm{d}x' \\
&= \frac{1}{V_{n,p}V_{n,p'}} \int_0^1 \int_0^1 \phi(x; p, \frac{pq}{2N})\phi(x'; p', \frac{p'q'}{2N}) \left[\sum_{i=1}^{\infty} \mathbb{P}_{\mathcal{E}_{n,x,x'}}(B_i) \right] \mathrm{d}x\mathrm{d}x' \\
&= \frac{1}{V_{n,p}V_{n,p'}} \int_0^1 \int_0^1 \phi(x; p, \frac{pq}{2N})\phi(x'; p', \frac{p'q'}{2N}) \mathbb{P}_{\mathcal{E}_{n,x,x'}}(\biguplus_{i=1}^{\infty} B_i) \mathrm{d}x\mathrm{d}x'
\end{aligned}$$

and $\biguplus_{i=1}^{\infty} B_i \in \mathcal{G}$. Since $\mathcal{G}$ is a λ-system and contains all rectangles, by Dynkin's theorem, it must also contain the σ-algebra generated by the rectangles, which is the discrete σ-algebra over E_n.

Statement 3 Take $d \times d' \in E_n^2$ satisfying both $|M(d) - 2Np| \leq 2y\sqrt{Npq}$ and $|M(d') - 2Np'| \leq 2y\sqrt{Np'q'}$. Then, using Theorem 4.20(3) we can write

$$\begin{aligned}
&\mathbb{P}_{\mathcal{I}_{n,p,p'}}(d \times d') \\
&= \mathbb{P}_{\mathcal{I}_{n,p}}(d) \cdot \mathbb{P}_{\mathcal{I}_{n,p}}(d') \\
&= \mathbb{P}_{\mathcal{E}'_{n,p}}(d) \left(1 + O\left(\frac{1+|y|^3}{\sqrt{pqN}}\right)\right) \mathbb{P}_{\mathcal{E}'_{n,p'}}(d') \left(1 + O\left(\frac{1+|y|^3}{\sqrt{p'q'N}}\right)\right).
\end{aligned}$$

Note that the inequality from Theorem 4.20(3) was applied twice, for $\mathcal{I}_{n,p}$ and for $\mathcal{I}_{n,p'}$. Since both inequalities are uniform in their respective sets — $\{d \in E_n : |M(d) - 2Np| \leq 2y\sqrt{Npq}\}$ and $\{d' \in E_n : |M(d') - 2Np'| \leq 2y\sqrt{Np'q'}\}$ —, the resulting inequality is uniform in the set $\{d \times d' \in E_n^2 : |M(d) - 2Np| \leq 2y\sqrt{Npq}, |M(d') - 2Np'| \leq 2y\sqrt{Np'q'}\}$. Algebraic manipulations yield

$$\begin{aligned}
\mathbb{P}_{\mathcal{I}_{n,p,p'}}(d \times d') = \mathbb{P}_{\mathcal{E}'_{n,p}}(d)\mathbb{P}_{\mathcal{E}'_{n,p'}}(d') &\left(1 + O\left(\frac{1+|y|^3}{\sqrt{pqN}}\right) + O\left(\frac{1+|y'|^3}{\sqrt{p'q'N}}\right)\right. \\
&\left. + O\left(\frac{1+|y|^3}{\sqrt{pqN}}\right) O\left(\frac{1+|y'|^3}{\sqrt{p'q'N}}\right)\right) \\
= \mathbb{P}_{\mathcal{E}'_{n,p,p'}}(d \times d') &\left(1 + O\left(\frac{1+|y|^3}{\sqrt{pqN}}\right) + O\left(\frac{1+|y|^3}{\sqrt{p'q'N}}\right)\right),
\end{aligned}$$

since $y = o(\sqrt[6]{pqN})$ (which implies $(1 + |y|^3)/\sqrt{pqN} = o(1)$). This proves the result.

Statement 4 This proof will follow by construction. Under the stated assumptions for p, there exist sets $R_p(n), R'_p(n) \subseteq E_n$ and a real function $\delta(n)$ satisfying the conditions of Theorem 4.20(4), with δ_d for each $d \in R_p(n)$ in condition (b). Applying the same reasoning for p', there are also sets $R_{p'}(n), R'_{p'}(n) \subseteq E_n$ and real function $\delta'(n)$ satisfying these same conditions, this time with $\delta'_{d'}$ for each $d' \in R_{p'}(n)$ in condition (b) (δ and δ' will not necessarily be equal for $d = d'$). Note that $\delta(n), \delta'(n)$ must be positive real functions.

Now, take:

$$\begin{aligned}
S_{p,p'}(n) &= R_p(n) \times R_{p'}(n), \\
S'_{p,p'}(n) &= R'_p(n) \times R'_{p'}(n), \\
\varepsilon(n) &= \delta(n) + \delta'(n).
\end{aligned}$$

We will show the desired results hold for $S_{p,p'}$, $S'_{p,p'}$ and ε, using properties of $R_p, R'_p, R_{p'}, R'_{p'}, \delta$, and δ' thoroughly in the next steps:

4a. Note that

$$\begin{aligned}\mathbb{P}_{\mathcal{D}_{n,p,p'}}(S_{p,p'}(n)) &= \mathbb{P}_{\mathcal{D}_{n,p}}(R_p(n)) \cdot \mathbb{P}_{\mathcal{D}_{n,p'}}(R_{p'}(n)) \\ &= (1 - n^{-\omega(n)})(1 - n^{-\omega(n)}) \\ &= 1 - n^{-\omega(n)},\end{aligned}$$

and similarly for $\mathbb{P}_{\mathcal{D}_{n,p,p'}}(S'_{p,p'}(n))$.

4b. Note that, for any $d \times d' \in S'_{p,p'}$, it holds that $d \in R'_p(n)$ and $d' \in R'_{p'}(n)$. The former implies $\gamma_2 = \lambda(1-\lambda)(1+o(1))$ and the latter implies $\gamma'_2 = \lambda'(1-\lambda')(1+o(1))$.

4c. By construction, for any $d \times d' \in S_{p,p'}(n)$, it holds that $d \in R_p(n)$ and $d' \in R_{p'}(n)$. Therefore, taking $\varepsilon_{d\times d'} = \delta_d + \delta'_{d'}$, it holds that $|\varepsilon_{d\times d'}| \leq |\delta_d| + |\delta'_{d'}| \leq \delta(n) + \delta'(n) = \varepsilon(n)$, and

$$\begin{aligned}&\frac{\mathbb{P}_{\mathcal{D}_{n,p,p'}}(d \times d')}{\mathbb{P}_{\mathcal{E}'_{n,p,p'}}(d \times d')} \\ &= \frac{\mathbb{P}_{\mathcal{D}_{n,p}}(d) \cdot \mathbb{P}_{\mathcal{D}_{n,p'}}(d')}{\mathbb{P}_{\mathcal{E}'_{n,p}}(d) \cdot \mathbb{P}_{\mathcal{E}'_{n,p'}}(d')} \\ &= \exp\left\{\frac{1}{4}\left(1 - \frac{\gamma_2^2}{\lambda^2(1-\lambda)^2}\right)\right\} \cdot \exp\{\delta_d\} \\ &\qquad \cdot \exp\left\{\frac{1}{4}\left(1 - \frac{(\gamma'_2)^2}{(\lambda')^2(1-\lambda')^2}\right)\right\} \cdot \exp\{\delta_{d'}\} \\ &= \exp\left\{\frac{1}{4}\left(2 - \frac{\gamma_2^2}{\lambda^2(1-\lambda)^2} - \frac{(\gamma'_2)^2}{(\lambda')^2(1-\lambda')^2}\right)\right\} \cdot \exp\{\delta_d + \delta_{d'}\} \\ &= \exp\left\{\frac{1}{4}\left(2 - \frac{\gamma_2^2}{\lambda^2(1-\lambda)^2} - \frac{(\gamma'_2)^2}{(\lambda')^2(1-\lambda')^2}\right)\right\} \cdot \exp\{\varepsilon_{d\times d'}\}.\end{aligned}$$

□

Using the above stepwise approximation through the models, we can derive a general-purpose rule for vanishing probabilities of events involving independent $G(n,p)$ random graphs, similar to the one stated in Theorem 4.21.

Theorem 4.23. *Let A_n be a sequence of events in E_n^2. If p, p' satisfy $\omega(\log n/n) \leq p, p' \leq o(n^{-1/2})$, then $\mathbb{P}_{\mathcal{B}_{n,p,p'}}(A_n) = o(n^{-a})$ implies $\mathbb{P}_{\mathcal{D}_{n,p,p'}}(A_n) = o(n^{-a})$ for any fixed $a > 0$.*

Proof. Before anything, we note that our hypotheses imply that $\omega(1/n) \leq p, p' \leq o(1)$, which we will use several times along the proof. Let $a > 0$ be fixed, and assume $\mathbb{P}_{\mathcal{B}_{n,p,p'}}(A_n) = o(n^{-a})$.

In agreement with the approximation scheme previously presented, we will prove our assertion in four steps, each addressing one of the following statements:

1. $\mathbb{P}_{\mathcal{B}_{n,p,p'}}(A_n) = o(n^{-a})$ implies $\mathbb{P}_{\mathcal{E}_{n,p,p'}}(A_n) = o(n^{-a})$;
2. $\mathbb{P}_{\mathcal{E}_{n,p,p'}}(A_n) = o(n^{-a})$ implies $\mathbb{P}_{\mathcal{I}_{n,p,p'}}(A_n) = o(n^{-a})$;
3. $\mathbb{P}_{\mathcal{I}_{n,p,p'}}(A_n) = o(n^{-a})$ implies $\mathbb{P}_{\mathcal{E}'_{n,p,p'}}(A_n) = o(n^{-a})$;
4. $\mathbb{P}_{\mathcal{E}'_{n,p,p'}}(A_n) = o(n^{-a})$ implies $\mathbb{P}_{\mathcal{D}_{n,p,p'}}(A_n) = o(n^{-a})$.

Step 1 Assume $\mathbb{P}_{\mathcal{B}_{n,p,p'}}(A_n) = o(n^{-a})$. Theorem 4.22(1) states that

$$\mathbb{P}_{\mathcal{E}_{n,p,p'}}(A_n) = \frac{4\mathbb{P}_{\mathcal{B}_{n,p,p'}}(A_n)}{[1+(q-p)^{2N}][1+(q'-p')^{2N}]}.$$

Since $p = \omega(1/n)$, it follows that $2Np \to \infty$ and $(q-p)^{2N} = (1 - 2Np/2N)^{2N} \to 0$. By an analogous argument, $(q'-p')^{2N} \to 0$. Thus, $\mathbb{P}_{\mathcal{E}_{n,p,p'}}(A_n) \sim 4\mathbb{P}_{\mathcal{B}_{n,p,p'}}(A_n)$ and, since $\mathbb{P}_{\mathcal{B}_{n,p,p'}}(A_n) = o(n^{-a})$, it follows that $\mathbb{P}_{\mathcal{E}_{n,p,p'}}(A_n) = o(n^{-a})$.

Step 2 Assume $\mathbb{P}_{\mathcal{E}_{n,p,p'}}(A_n) = o(n^{-a})$. We turn to the expression that links $\mathcal{E}_{n,p,p'}$ to $\mathcal{I}_{n,p,p'}$, presented in Theorem 4.22(2).

The normalization constant $V_{n,p}V_{n,p'}$ is the probability that two independent random variables $\mathrm{N}(p, pq/2N)$ and $\mathrm{N}(p', p'q'/2N)$ assume values in $[0,1]$. Standardizing these random variables and denoting by $Q(\cdot)$ the Q-function (tail distribution of a standard normal random variable), we have

$$\begin{aligned} V_{n,p} &= Q\left(-\frac{p}{\sqrt{pq/2N}}\right) - Q\left(\frac{q}{\sqrt{pq/2N}}\right) \\ &= Q\left(-\sqrt{\frac{2Np}{q}}\right) - Q\left(\sqrt{\frac{2Nq}{p}}\right) \\ &\to 1, \end{aligned}$$

where the limit comes from the facts that $2Np/q = \omega(1)$ whenever $p = \omega(1/n)$ and $2Nq/p = \omega(1)$ whenever $p = o(1)$. The same limit applies to $V_{n,p'}$ by the same argument. Thus $1/V_{n,p}V_{n,p'} = \theta(1)$.

For the integral, we will split the domain of integration into several rectangles and deal with them separately. To simplify our notation, we denote our integrand by $g(x,x') = \phi(x; p, \frac{pq}{2N})\phi(x'; p', \frac{p'q'}{2N})\mathbb{P}_{\mathcal{E}_{n,x,x'}}(A_n)$. Pick some constant $c > a$, and let $\delta = \delta(n) = c\sqrt{q\log n/Np}$ and $\delta' = \delta'(n) = c\sqrt{q'\log n/Np'}$. Note that $np = \omega(\log n)$ implies $\delta = c\sqrt{pq\log n/Np^2} = o(pq\log n/\log^2 n) = o(p)$. Similarly, $np' = \omega(\log n)$ implies $\delta' = o(p')$. Since $p = o(q)$ whenever $p = o(1)$, it holds that $\delta < p, q$ for large enough n. For such n, we can safely

write

$$\begin{aligned}
&\int_0^1 \int_0^1 \phi(x; p, \frac{pq}{2N})\phi(x'; p', \frac{p'q'}{2N})\mathbb{P}_{\mathcal{E}_{n,x,x'}}(A_n)\mathrm{d}x\mathrm{d}x' \\
&= \int_0^1 \int_0^1 g(x, x')\mathrm{d}x\mathrm{d}x' \\
&= \iint_{(1)} g(x, x')\mathrm{d}x\mathrm{d}x' + \iint_{(2)} g(x, x')\mathrm{d}x\mathrm{d}x' + \iint_{(3)} g(x, x')\mathrm{d}x\mathrm{d}x' \\
&+ \iint_{(4)} g(x, x')\mathrm{d}x\mathrm{d}x' + \iint_{(5)} g(x, x')\mathrm{d}x\mathrm{d}x' + \iint_{(6)} g(x, x')\mathrm{d}x\mathrm{d}x' \\
&+ \iint_{(7)} g(x, x')\mathrm{d}x\mathrm{d}x' + \iint_{(8)} g(x, x')\mathrm{d}x\mathrm{d}x' + \iint_{(9)} g(x, x')\mathrm{d}x\mathrm{d}x',
\end{aligned}$$

where the subdomains of integration are as follows:

$$\begin{aligned}
(1) &= [0, p(1-\delta)) \times [0, p'(1-\delta')), \\
(2) &= [0, p(1-\delta)) \times [p'(1-\delta'), p'(1+\delta')], \\
(3) &= [0, p(1-\delta)) \times (p'(1+\delta'), 1], \\
(4) &= [p(1-\delta), p(1+\delta)] \times [0, p'(1-\delta')), \\
(5) &= [p(1-\delta), p(1+\delta)] \times [p'(1-\delta'), p'(1+\delta')], \\
(6) &= [p(1-\delta), p(1+\delta)] \times (p'(1+\delta'), 1], \\
(7) &= (p(1+\delta), 1] \times [0, p'(1-\delta')), \\
(8) &= (p(1+\delta), 1] \times [p'(1-\delta'), p'(1+\delta')], \\
(9) &= (p(1+\delta), 1] \times (p'(1+\delta'), 1].
\end{aligned}$$

These subdomains are illustrated in Figure 4.1.

The integral over the corner subdomain (1) can be bounded easily, by noting that $\mathbb{P}_{\mathcal{E}_{n,x,x'}}(A_n) \leq 1$:

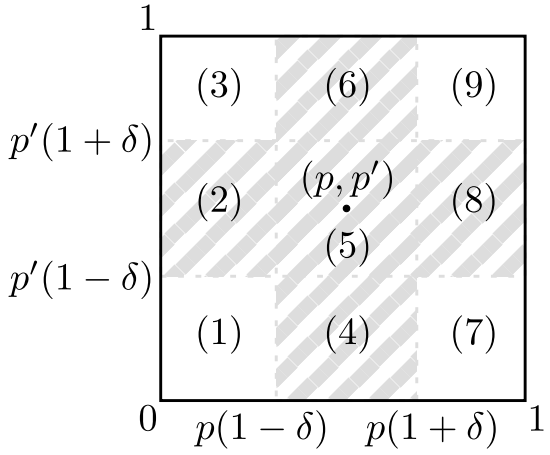


Figure 4.1: Splitting $[0,1] \times [0,1]$ into 9 smaller domains of integration. We call domains (1), (3), (7), and (9) the *corner* subdomains, domains (2), (4), (6), and (8) the *side* subdomains, and domain (5) the *central* subdomain.

$$
\begin{aligned}
&\int_0^{p(1-\delta)} \int_0^{p(1-\delta')} g(x,x')\mathrm{d}x\mathrm{d}x' \\
&= \int_0^{p(1-\delta)} \int_0^{p(1-\delta')} \phi(x;p,\frac{pq}{2N})\phi(x';p',\frac{p'q'}{2N})\mathbb{P}_{\mathcal{E}_{n,x,x'}}(A_n)\mathrm{d}x\mathrm{d}x' \\
&\leq \int_0^{p(1-\delta)} \int_0^{p(1-\delta')} \phi(x;p,\frac{pq}{2N})\phi(x';p',\frac{p'q'}{2N})\mathrm{d}x\mathrm{d}x' \\
&= \left[\int_0^{p(1-\delta)} \phi(x;p,\frac{pq}{2N})\mathrm{d}x\right]\left[\int_0^{p(1-\delta')} \phi(x';p',\frac{p'q'}{2N})\mathrm{d}x'\right] \\
&\leq Q\left(\frac{p\delta}{\sqrt{pq/2N}}\right) Q\left(\frac{p'\delta'}{\sqrt{p'q'/2N}}\right) \leq \exp\left\{-\frac{Np\delta^2}{q}\right\}\exp\left\{-\frac{Np'(\delta')^2}{q'}\right\} \\
&= \exp\{-c\log n\}\exp\{-c\log n\} = n^{-2c} = o(n^{-a}),
\end{aligned}
$$

where the last step comes from the choice of c. Similarly we assert this bound for subdomains (3), (7), and (9), with a few changes in the arguments of the Q-functions.

For the integral on side subdomain (2), we can follow a similar strategy, but a little bit more carefully:

$$
\begin{aligned}
&\int_0^{p(1-\delta)} \int_{p(1-\delta')}^{p(1+\delta')} g(x,x')\mathrm{d}x\mathrm{d}x' \\
&= \int_0^{p(1-\delta)} \int_{p(1-\delta')}^{p(1+\delta')} \phi(x;p,\frac{pq}{2N})\phi(x';p',\frac{p'q'}{2N})\mathbb{P}_{\mathcal{E}_{n,x,x'}}(A_n)\mathrm{d}x\mathrm{d}x' \\
&\leq \int_0^{p(1-\delta)} \int_{p(1-\delta')}^{p(1+\delta')} \phi(x;p,\frac{pq}{2N})\phi(x';p',\frac{p'q'}{2N})\mathrm{d}x\mathrm{d}x' \\
&= \left[\int_0^{p(1-\delta)} \phi(x;p,\frac{pq}{2N})\mathrm{d}x\right]\left[\int_{p(1-\delta')}^{p(1+\delta')} \phi(x';p',\frac{p'q'}{2N})\mathrm{d}x'\right] \\
&\leq Q\left(\frac{p\delta}{\sqrt{pq/2N}}\right)\cdot 1 \leq \exp\left\{-\frac{Np\delta^2}{q}\right\} \\
&= \exp\{-c\log n\} = n^{-c} = o(n^{-a}).
\end{aligned}
$$

Again, the same bound holds for subdomains (4), (6), and (8), with minor changes in the arguments of the Q-functions.

For the integral on center subdomain (5), some comments are appropriate. First, note that, since $\delta = o(p)$, for any $x = x(n) \in [p(1-\delta), p(1+\delta)]$ it is true that $x = p(1+o(p))$ and, therefore, x has the same asymptotics as p — namely, $o(n^{-1/2}) \leq x \leq \omega(\log n/n)$. Similarly, for any $x' = x'(n) \in [p'(1-\delta'), p'(1+\delta')]$, it holds that $x' = o(n^{-1/2})$ and $x' = \omega(\log n/n)$.

Also, for any fixed n, $\mathbb{P}_{\mathcal{E}_{n,p,p'}}(A_n)$ is a continuous function of p and p'. This comes from the result of Theorem 4.22(1) and the fact that $\mathbb{P}_{\mathcal{B}_{n,p,p'}}(A_n) = \sum_{d\times d'\in A_n} \mathbb{P}_{\mathcal{B}_{n,p,p'}}(d\times d')$. Since the probability of each such $d\times d'$ under the measure $\mathbb{P}_{\mathcal{B}_{n,p,p'}}$ is a continuous function of p and p' (product of powers of p, p', $1-p$ and $1-p'$ and some constants in both p and p'), and the sum of these functions has a finite number of terms, continuity of $\mathbb{P}_{\mathcal{B}_{n,p,p'}}(A_n)$ with respect to p and p' follows. Since, by Theorem 4.22(1), $\mathbb{P}_{\mathcal{E}_{n,p,p'}}(A_n)$ is the product between $\mathbb{P}_{\mathcal{B}_{n,p,p'}}(A_n)$ and a continuous function of p and p', continuity of the former with respect to p and p' also follows.

As a consequence of these results, for $(x,x') \in [p(1-\delta), p(1+\delta)] \times [p'(1-\delta'), p'(1+\delta')]$, the function $\mathbb{P}_{\mathcal{E}_{n,x,x'}}(A_n)$, being a continuous function over this compact set, will attain a maximum value for some argument $(y,y')(n)$ in this set. Such (y,y') will, forcefully, satisfy $y, y' = o(n^{-1/2})$ and $y, y' = \omega(\log n/n)$, which means, by our conclusion from the previous step, that $\mathbb{P}_{\mathcal{E}_{n,y,y'}}(A_n) = o(n^{-a})$.

That being said, we can assert that

$$
\begin{aligned}
&\int_{p(1-\delta)}^{p(1+\delta)} \int_{p'(1-\delta')}^{p'(1+\delta')} g(x,x')\mathrm{d}x\mathrm{d}x' \\
&= \int_{p(1-\delta)}^{p(1+\delta)} \int_{p'(1-\delta')}^{p'(1+\delta')} \phi(x;p,\frac{pq}{2N})\phi(x';p',\frac{p'q'}{2N})\mathbb{P}_{\mathcal{E}_{n,x,x'}}(A_n)\mathrm{d}x\mathrm{d}x' \\
&\leq \int_{p(1-\delta)}^{p(1+\delta)} \int_{p'(1-\delta')}^{p'(1+\delta')} \phi(x;p,\frac{pq}{2N})\phi(x';p',\frac{p'q'}{2N}) \cdot \\
&\qquad \left[\max_{(x,x')\in[p(1-\delta),p(1+\delta)]\times[p'(1-\delta'),p'(1+\delta')]} \mathbb{P}_{\mathcal{E}_{n,x,x'}}(A_n)\right]\mathrm{d}x\mathrm{d}x' \\
&= \int_{p(1-\delta)}^{p(1+\delta)} \int_{p'(1-\delta')}^{p'(1+\delta')} \phi(x;p,\frac{pq}{2N})\phi(x';p',\frac{p'q'}{2N})\mathbb{P}_{\mathcal{E}_{n,y,y'}}(A_n)\mathrm{d}x\mathrm{d}x' \\
&= \mathbb{P}_{\mathcal{E}_{n,y,y'}}(A_n) \int_{p(1-\delta)}^{p(1+\delta)} \int_{p'(1-\delta')}^{p'(1+\delta')} \phi(x;p,\frac{pq}{2N})\phi(x';p',\frac{p'q'}{2N})\mathrm{d}x\mathrm{d}x' \\
&\leq \mathbb{P}_{\mathcal{E}_{n,y,y'}}(A_n) \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \phi(x;p,\frac{pq}{2N})\phi(x';p',\frac{p'q'}{2N})\mathrm{d}x\mathrm{d}x' \\
&= o(n^{-a}) \cdot 1 = o(n^{-a}).
\end{aligned}
$$

Thus, we conclude that

$$\mathbb{P}_{\mathcal{I}_{n,p,p'}}(A_n) = \theta(1) \cdot (9 \cdot o(n^{-a})) = o(n^{-a}).$$

Step 3 Assume $\mathbb{P}_{\mathcal{I}_{n,p,p'}}(A_n) = o(n^{-a})$. We begin by recalling that

$$\left(\frac{1}{2}M(S_1), \frac{1}{2}M(S_2)\right) \overset{d}{\sim} \mathrm{Bin}(N,p) \otimes \mathrm{Bin}(N,p') \text{ under } \mathbb{P}_{\mathcal{E}'_{n,p,p'}}.$$

Define the event $N_n = \{|M(S_1) - 2Np| < 2Np \cdot \varepsilon, |M(S_2) - 2Np'| < 2Np' \cdot \varepsilon'\}$, with $\varepsilon = (2Np)^{-5/12}$ and $\varepsilon' = (2Np')^{-5/12}$. By the Chernoff bound, we have

$$
\begin{aligned}
\mathbb{P}_{\mathcal{E}'_{n,p,p'}}(|M(S_1) - 2Np| \geq 2Np \cdot \varepsilon) &\leq 2e^{-2Np\varepsilon^2/6} = 2e^{-\frac{1}{6}(2Np)^{1/6}}, \\
\mathbb{P}_{\mathcal{E}'_{n,p,p'}}(|M(S_2) - 2Np'| \geq 2Np' \cdot \varepsilon) &\leq 2e^{-2Np'(\varepsilon')^2/6} = 2e^{-\frac{1}{6}(2Np')^{1/6}}.
\end{aligned}
$$

Thus, by the union bound, $\mathbb{P}_{\mathcal{E}'_{n,p,p}}(\overline{N_n}) \leq 2(e^{-\frac{1}{6}(2Np)^{1/6}} + e^{-\frac{1}{6}(2Np')^{1/6}})$.

Now, by the definition of the event N_n, it holds that, in this event:

$$\begin{aligned}|M(S_1) - 2Np| < 2Np \cdot \varepsilon &= (2Np)(2Np)^{-5/12}\frac{\sqrt{4Npq}}{\sqrt{4Npq}} \\ &= \left[\frac{(2Np)^{1/12}}{\sqrt{2q}}\right]\sqrt{4Npq}, \\ |M(S_2) - 2Np'| < 2Np' \cdot \varepsilon' &= (2Np')(2Np')^{-5/12}\frac{\sqrt{4Np'q'}}{\sqrt{4Np'q'}} \\ &= \left[\frac{(2Np')^{1/12}}{\sqrt{2q'}}\right]\sqrt{4Np'q'}.\end{aligned}$$

The same inequalities hold in the event $A_n \cap N_n \subseteq N_n$. Now, note that $(2Np)^{1/12}/\sqrt{2q} = o(\sqrt[6]{2Npq})$ and $(2Np')^{1/12}/\sqrt{2q'} = o(\sqrt[6]{2Np'q'})$, which allows us to relate the probability of $A_n \cap N_n$ under measures $\mathbb{P}_{\mathcal{E}'_{n,p,p'}}$ and $\mathbb{P}_{\mathcal{I}_{n,p,p'}}$. We choose $y = \max\{(2Np)^{1/12}/\sqrt{2q}, (2Np')^{1/12}/\sqrt{2q'}\}$; this choice of y and $q = \theta(1)$ imply $(1+|y|^3)/\sqrt{pqN} = o((Np)^{-1/2}) + o(n^{3/8})/\omega(n^{1/2}) = o(1)$ and $(1+|y|^3)/\sqrt{p'qN} = o((Np')^{-1/2}) + o(n^{3/8})/\omega(n^{1/2}) = o(1)$. From these facts, using Theorem 4.22(3), it follows that

$$\begin{aligned}\mathbb{P}_{\mathcal{E}'_{n,p,p'}}(A_n) &= \mathbb{P}_{\mathcal{E}'_{n,p,p'}}(A_n \cap \overline{N_n}) + \mathbb{P}_{\mathcal{E}'_{n,p,p'}}(A_n \cap N_n) \\ &\leq \mathbb{P}_{\mathcal{E}'_{n,p,p'}}(\overline{N_n}) + \mathbb{P}_{\mathcal{E}'_{n,p,p'}}(A_n \cap N_n) \\ &\leq 2(e^{-\frac{1}{6}(2Np)^{1/6}} + e^{-\frac{1}{6}(2Np')^{1/6}}) \\ &\quad + \mathbb{P}_{\mathcal{I}_{n,p,p'}}(A_n \cap \overline{N_n})\left(1 + O\left(\frac{1+|y|^3}{\sqrt{pqN}}\right) + O\left(\frac{1+|y|^3}{\sqrt{p'q'N}}\right)\right)^{-1} \\ &\leq e^{-\omega(n)} + \mathbb{P}_{\mathcal{I}_{n,p,p'}}(A_n)(1 + o(1) + o(1))^{-1} \\ &= o(n^{-a}) + o(n^{-a})(\theta(1))^{-1} = o(n^{-a}).\end{aligned}$$

Step 4 Assume $\mathbb{P}_{\mathcal{E}'_{n,p,p'}}(A_n) = o(n^{-a})$. Let the sets $S_{p,p'}(n), S'_{p,p'}(n)$ and the real function $\varepsilon(n)$ be as in Theorem 4.22(4) (note that our hypotheses about p, p' imply the hypotheses of this theorem are satisfied), and define the set $T_{p,p'}(n) = S_{p,p'}(n) \cap S'_{p,p'}(n)$. Then the following facts hold:

1. $\mathbb{P}_{\mathcal{D}_{n,p,p'}}(T_{p,p'}(n)) = 1 - n^{-\omega(n)}$ (as, by union bound, $\mathbb{P}_{\mathcal{D}_{n,p,p'}}(\overline{T_{p,p'}(n)}) \leq \mathbb{P}_{\mathcal{D}_{n,p,p'}}(\overline{S_{p,p'}(n)}) + \mathbb{P}_{\mathcal{D}_{n,p,p'}}(\overline{S'_{p,p'}(n)}) = 2n^{-\omega(n)} = n^{-\omega(n)}$);
2. for every $d \times d' \in T_{p,p'}(n)$, there is some $\varepsilon_{d\times d'}$ such that $|\varepsilon_{d\times d'}| \leq \varepsilon(n)$ and

$$\frac{\mathbb{P}_{\mathcal{D}_{n,p,p'}}(d \times d')}{\mathbb{P}_{\mathcal{E}'_{n,p,p'}}(d \times d')} = \exp\left\{\frac{1}{4}\left(2 - \frac{\gamma_2^2}{\lambda^2(1-\lambda)^2} - \frac{(\gamma_2')^2}{(\lambda')^2(1-\lambda')^2}\right)\right\} \cdot \exp\{\varepsilon_{d\times d'}\};$$

3. in $T_{p,p'}(n)$, $\gamma_2 = \lambda(1-\lambda)(1+o(1))$ and $\gamma_2' = \lambda'(1-\lambda')(1+o(1))$;

Using these facts, it follows that:

$$\begin{aligned}
\mathbb{P}_{\mathcal{D}_{n,p,p'}}(A_n) &= \mathbb{P}_{\mathcal{D}_{n,p,p'}}(A_n \cap \overline{T_{p,p'}(n)}) + \mathbb{P}_{\mathcal{D}_{n,p,p'}}(A_n \cap T_{p,p'}(n)) \\
&\leq \mathbb{P}_{\mathcal{D}_{n,p,p'}}(\overline{T_{p,p'}(n)}) + \sum_{d\times d' \in A_n \cap T_{p,p'}(n)} \mathbb{P}_{\mathcal{D}_{n,p,p'}}(d \times d') \\
&= n^{-\omega(n)} + \sum_{d\times d' \in A_n \cap T_{p,p'}(n)} \Bigg[\mathbb{P}_{\mathcal{E}'_{n,p,p'}}(d \times d') \cdot \\
&\qquad \exp\left\{ \frac{1}{4} \left(2 - \frac{\gamma_2^2}{\lambda^2(1-\lambda)^2} - \frac{(\gamma_2')^2}{(\lambda')^2(1-\lambda')^2} \right) \right\} \cdot \exp\{\varepsilon_{d\times d'}\} \Bigg] \\
&= n^{-\omega(n)} + \left[\sum_{d\times d' \in A_n \cap T_{p,p'}(n)} \mathbb{P}_{\mathcal{E}'_{n,p,p'}}(d \times d') \right] \cdot \\
&\qquad \max_{d\times d' \in A_n \cap T_{p,p'}(n)} \exp\left\{ \frac{1}{4} \left(2 - \frac{\gamma_2^2}{\lambda^2(1-\lambda)^2} - \frac{(\gamma_2')^2}{(\lambda')^2(1-\lambda')^2} \right) \right\} \cdot \\
&\qquad \max_{d\times d' \in A_n \cap T_{p,p'}(n)} \exp\{\varepsilon_{d\times d'}\} \\
&\leq n^{-\omega(n)} + \mathbb{P}_{\mathcal{E}'_{n,p,p'}}(A_n \cap T_{p,p'}(n)) \cdot \\
&\qquad \exp\left\{ \frac{1}{4}(2 - (1+o(1))^2 - (1+o(1))^2) \right\} \cdot \exp\{\varepsilon(n)\} \\
&\leq o(n^{-a}) + \mathbb{P}_{\mathcal{E}'_{n,p,p'}}(A_n) \cdot \exp\{o(1)\} \cdot \exp\{o(1)\} \\
&= o(n^{-a}) + o(n^{-a}) \cdot \theta(1) \cdot \theta(1) = o(n^{-a}).
\end{aligned}$$

□

4.3.3 Degree sequences in $G(n,p)$

Theorem 4.22 allows us to handle the degree sequences of two $G(n,p)$ random graphs as sequences of independent random variables. Our goal is to ensure that the edit distance Δ between these degree sequences is sufficiently high, with high probability, thus executing step 5 of our main proof. This is achieved by the following theorem:

Theorem 4.24. *Let G_1 and G_2 be independent $G(n,p)$ random graphs, with p satisfying $\omega(\log n/n) \leq p \leq o(n^{-1/2})$. Then, for any $\varepsilon > 0$ and any $a > 0$,*

$$\Delta(G_1, G_2) \geq n^{1/2-\varepsilon}$$

with probability $1 - o(n^{-a})$.

We remark that this theorem will be later applied with G_1 and G_2 being independent 1-neighborhoods of two vertices, say v_1 and v_2, in a $G(n,p)$. In this context,

the degree sequences taken in the calculation of Δ are interpreted as "local degrees" of neighbors of v_1 and v_2, that is, degrees of neighbors of v_1 and v_2 restricted to within $\mathcal{N}[v_1]$ and $\mathcal{N}[v_2]$, respectively. However, this statement is more general in that it takes as input two independent $G(n,p)$ random graphs, regardless of how they were obtained, so the degree sequences involved are the degrees of all vertices in G_1 and G_2. We still need some additional intermediate results before proceeding to its proof.

Lemma 4.25. *Let X be a $\mathrm{Bin}(n,p)$ random variable with $p < 1/2$. If $\varepsilon > 0$ and $np\varepsilon^2 \geq 3$, then*

$$\mathbb{P}[|X - np| \geq \varepsilon np] \geq 2\exp\{-9np\varepsilon^2\}.$$

Proof. See Lemma 5.2 of [38]. □

Lemma 4.26. *Let $\vec{X}$ be a multinomial random vector $\mathrm{Mult}(n, p_1, \ldots, p_k)$ with k fixed, and let $0 < \beta < 1$ be also fixed. If $\Omega(n^{-\beta}) \leq p_1, \ldots, p_{k-1} \leq o(1)$, then*

$$\max_{\vec{x}} \mathbb{P}[\vec{X} = \vec{x}] = O(n^{-(k-1)(1-\beta)/2}).$$

Proof. Let $x^* = (x_1^*, \ldots, x_k^*)$ be the mode of $\mathrm{Mult}(n, p_1, \ldots, p_k)$. It is known [39] that $x_i^* = I(np_i)$, where $I(a)$ is either $\lfloor a \rfloor$ or $\lceil a-1 \rceil$. This implies that $x_i^* \leq np_i \leq x_i^* + 1$. Also, since $\beta < 1$, it holds that $np_i \to \infty$ for all i, which implies that, for large enough n, $x_i^* \geq 1$ for all i.

Using these inequalities, Stirling's approximation, and known bounds for the exponential function, we have, for large enough n:

$$\begin{aligned}
\max_{\vec{x}} \mathbb{P}[\vec{X} = \vec{x}] = \mathbb{P}[\vec{X} = x^*] &= \frac{n!}{\prod_{i=1}^n x_i^*!} \prod_{i=1}^n p_i^{x_i^*} \\
&\leq \frac{en^n e^{-n}\sqrt{n}}{(\sqrt{2\pi})^k \prod_{i=1}^k (x_i^*)^{x_i^*} \sqrt{x_i^*} e^{-x_i^*}} \prod_{i=1}^k p_i^{x_i^*} \\
&\leq \frac{e\sqrt{n}}{(\sqrt{2\pi})^k} \prod_{i=1}^k \left(\frac{np_i}{x_i^*}\right)^{x_i^*} \frac{1}{\sqrt{x_i^* \cdots x_{k-1}^*}} \frac{1}{\sqrt{np_k}} \\
&\leq \frac{e}{(\sqrt{2\pi})^k} \prod_{i=1}^k \left(1 + \frac{1}{x_i^*}\right) \frac{1}{\sqrt{np_1 \cdots np_{k-1}}} \frac{1}{\sqrt{p_k}} \\
&\leq \frac{e}{(\sqrt{2\pi})^k} \cdot e^k \cdot \frac{1}{\sqrt{np_1 \cdots np_{k-1}}} \frac{1}{\sqrt{p_k}}.
\end{aligned}$$

Now, the following inequalities hold for large enough n. First, by hypothesis, for every $i \leq k-1$, we have $p_i \geq c_i n^{-\beta}$ for some constant c_i. Second, $p_k \geq c_k$ for some constant c_k, since the hypotheses imply that $p_k \to 1$. Thus:

$$\begin{aligned}\max_{\vec{x}} \mathbb{P}[\vec{X} = \vec{x}] &\leq \frac{e}{(\sqrt{2\pi})^k} \cdot e^k \cdot \frac{1}{\sqrt{np_1 \cdots np_{k-1}}} \frac{1}{\sqrt{p_k}} \\ &\leq \frac{e^{k+1}}{(\sqrt{2\pi})^k} \frac{1}{\sqrt{c_1 \cdots c_{k-1}} \cdot c_k} \frac{1}{(\sqrt{n^{1-\beta}})^{k-1}} \\ &= K n^{-(k-1)(1-\beta)/2}\end{aligned}$$

for some constant K. This concludes the proof. □

This proof of Theorem 4.24 will proceed as follows. Instead of looking at the whole degree sequences of G_1 and G_2, we will group several ranges of degrees into "buckets", according to their distances to the expected degree of $G(n,p)$. This will allow us to bound the probability that each vertex belongs to each bucket, using Chernoff-like bounds. Considering the degrees as independent random variables characterizes the bucketed degree sequences as multinomial random vectors. Furthermore, the buckets themselves are carefully chosen so that the distribution of these vectors is not too concentrated, i.e., the probability of their modes is large enough. This means these two vectors will most likely be far apart from each other in L_1-norm, which is the desired result.

Proof of Theorem 4.24. We begin by noting that $\Delta(G_1, G_2) = \int |\phi_{G_1} - \phi_{G_2}| \, \mathrm{d}\mu$ is a function of the degree sequences of G_1 and G_2. Let ϕ'_{G_1}, ϕ'_{G_2} be the degree functions obtained by approximating these degree sequences by sequences of n independent $\mathrm{Bin}(n-1,p)$ random variables. By virtue of Theorem 4.17, it is enough to show that

$$\int |\phi'_{G_1} - \phi'_{G_2}| \, \mathrm{d}\mu \geq n^{1/2-\varepsilon}$$

with probability $1 - o(n^{-a})$.

Fix real positive numbers $\alpha > \sqrt{27}$ and $\beta < \varepsilon$, and choose a natural number $b > 2a/(\varepsilon - \beta)$ (the reason for the choice of α will be given later in this proof). Let the real positive intervals $S_1, \ldots, S_b$ be defined as

$$S_i = \left[\alpha^{-(i+1)} \sqrt{(n-1)p} \cdot f(n,p), \alpha^{-i} \sqrt{(n-1)p} \cdot f(n,p)\right),$$

where $f(n,p) = (\min\{\log n, (n-1)p\})^{1/4}$. To ease the notation, let the set $S_{b+1} = \mathbb{R}_+ \setminus \cup_{i=1}^{b} S_i$ contain the remainder of the positive real line.

We now use these sets to group the vertices in both graphs into "buckets",

according to their degrees, with set S_i indicating the set of allowed degrees according to their distances to the average degree $(n-1)p$. More formally, define the sets of integers $B_1, \ldots, B_{b+1}$ as

$$B_i = \{k \in \mathbb{N}_0 \ : \ |k - (n-1)p| \in S_i\}.$$

Note that $\cup_i B_i = \mathbb{N}_0$ and $B_i \cap B_j = \emptyset$ whenever $i \neq j$. Now, write $T_i^{(1)} = \int_{B_i} \phi'_{G_1} \, \mathrm{d}\mu$. $T_i^{(1)}$ counts the number of vertices of G_1 with degrees in B_i. Similarly, write $T_i^{(2)} = \int_{B_i} \phi'_{G_2} \, \mathrm{d}\mu$. Then it holds that

$$\begin{aligned}
\int |\phi'_{G_1} - \phi'_{G_2}| \, \mathrm{d}\mu &= \sum_{i=1}^{b+1} \int_{B_i} |\phi'_{G_1} - \phi'_{G_2}| \, \mathrm{d}\mu \\
&\geq \sum_{i=1}^{b+1} \left| \int_{B_i} \phi'_{G_1} \mathrm{d}\mu - \int_{B_i} \phi'_{G_2} \, \mathrm{d}\mu \right| \\
&= \sum_{i=1}^{b+1} |T_i^{(1)} - T_i^{(2)}| = \|\vec{T}^{(1)} - \vec{T}^{(2)}\|_1,
\end{aligned}$$

where $\vec{T}^{(j)} = (T_1^{(j)}, \ldots, T_{b+1}^{(j)})$.

This means that

$$\begin{aligned}
\mathbb{P}\left[\int |\phi'_{G_1} - \phi'_{G_2}| \, \mathrm{d}\mu \leq n^{1/2-\varepsilon}\right] &\leq \mathbb{P}\left[\|\vec{T}^{(1)} - \vec{T}^{(2)}\|_1 \leq n^{1/2-\varepsilon}\right] \\
&= \sum_{\vec{t}} \mathbb{P}\left[\vec{T}^{(1)} = \vec{t}, \|\vec{T}^{(2)} - \vec{t}\|_1 \leq n^{1/2-\varepsilon}\right] \\
&= \sum_{\vec{t}} \mathbb{P}\left[\vec{T}^{(1)} = \vec{t}\right] \mathbb{P}\left[\|\vec{T}^{(2)} - \vec{t}\|_1 \leq n^{1/2-\varepsilon}\right] \\
&\leq \sum_{\vec{t}} \mathbb{P}\left[\vec{T}^{(1)} = \vec{t}\right] \max_{\vec{t}} \mathbb{P}\left[\|\vec{T}^{(2)} - \vec{t}\|_1 \leq n^{1/2-\varepsilon}\right] \\
&= \max_{\vec{t}} \mathbb{P}\left[\|\vec{T}^{(2)} - \vec{t}\|_1 \leq n^{1/2-\varepsilon}\right].
\end{aligned}$$

Note that, for any $\vec{t}$, the event $\{\|\vec{T}^{(2)} - \vec{t}\|_1 \leq n^{1/2-\varepsilon}\}$ has at most $(2n^{1/2-\varepsilon} + 1)^b$ elements: each of the first b coordinates of $\vec{T}^{(2)}$ must be at distance at most $n^{1/2-\varepsilon}$ from the corresponding coordinate of $\vec{t}$ (for a maximum of $2n^{1/2-\varepsilon} + 1$ valid options), and the last one is uniquely determined from the previous choices, since the coordinates must sum up to n. This means that:

$$\begin{aligned}
\mathbb{P}\left[\int |\phi'_{G_1} - \phi'_{G_2}| \, \mathrm{d}\mu \leq n^{1/2-\varepsilon}\right] &\leq \max_{\vec{t}} \mathbb{P}\left[\|\vec{T}^{(2)} - \vec{t}\|_1 \leq n^{1/2-\varepsilon}\right] \\
&\leq (2n^{1/2-\varepsilon} + 1)^b \max_{\vec{t}} \mathbb{P}\left[\vec{T}^{(2)} = \vec{t}\right].
\end{aligned}$$

Now, the degree of every vertex in G_1 or G_2 must belong to some B_i, and these degrees are deemed to be i.i.d. random variables, by our initial argument regarding ϕ'_{G_1} and ϕ'_{G_2}. This implies that $\vec{T}^{(1)}, \vec{T}^{(2)}$ are multinomial random vectors $\mathrm{Mult}(n, p_1, \ldots, p_{b+1})$, where $p_i = \mathbb{P}[\mathrm{Bin}(n-1,p) \in B_i]$ is the probability that the degree of a vertex belongs to B_i.

At this moment, to apply Lemma 4.26, we would like to bound from both sides the value of p_i (for $i \leq b$). For an upper bound, an application of the Chernoff bound suffices:

$$\begin{aligned} p_i &= \mathbb{P}[\mathrm{Bin}(n-1,p) \in B_i] \\ &\leq \mathbb{P}[|\mathrm{Bin}(n-1,p) - (n-1)p| \geq \alpha^{-(i-1)}\sqrt{(n-1)p} \cdot f(n,p)] \\ &\leq 2\exp\{-f(n,p)^2\alpha^{-2(i+1)}/3\}, \end{aligned}$$

which is $o(1)$, since $f(n,p) \to \infty$.

For a lower bound, we use the Chernoff bound and an application of Lemma 4.25, noting that $\alpha^{-2(i+1)}f(n,p)^2 \geq 3$ for large enough n:

$$\begin{aligned} p_i &= \mathbb{P}[\mathrm{Bin}(n-1,p) \in B_i] \\ &= \mathbb{P}[|\mathrm{Bin}(n-1,p) - (n-1)p| \geq \alpha^{-(i-1)}\sqrt{(n-1)p} \cdot f(n,p)] \\ &\qquad - \mathbb{P}[|\mathrm{Bin}(n-1,p) - (n-1)p| \geq \alpha^{-i}\sqrt{(n-1)p} \cdot f(n,p)] \\ &\geq 2\exp\{-9\alpha^{-2(i+1)}f(n,p)^2\} - 2\exp\{-\alpha^{-2i}f(n,p)^2/3\} \\ &= 2\exp\{-9\alpha^{-2(i+1)}f(n,p)^2\}(1 - \exp\{-\alpha^{-2(i+1)}f(n,p)^2\}^{\gamma}), \end{aligned}$$

where, in the last passage, we let $\gamma = \alpha^2/3 - 9 > 0$. Note that $\gamma > 0$ and $f(n,p) \to \infty$ imply that the term inside the parentheses tends to 1. The remaining exponential satisfies

$$\begin{aligned} n^{\beta}\exp\{-9\alpha^{-2(i+1)}f(n,p)^2\} &\geq \exp\{\beta\log n - 9\alpha^{-2(i+1)}\sqrt{\log n}\} \\ &= \exp\{\beta\log n(1 - 9\alpha^{-2(i+1)}(\log^{-1/2} n))\} \\ &= \omega(1), \end{aligned}$$

thus the expression on the right-hand side is $\omega(n^{-\beta})$.

This means that the conditions of Lemma 4.26 are satisfied for random vectors $\vec{T}^{(1)}$ and $\vec{T}^{(2)}$, with $k = b+1$. Therefore, $\max_{\vec{t}} \mathbb{P}[\vec{T}^{(2)} = \vec{t}] = O(n^{-b(1-\beta)/2})$, and, by

the choices of b and β,

$$\begin{aligned}\mathbb{P}\left[\int |\phi'_{G_1} - \phi'_{G_2}| \, \mathrm{d}\mu \leq n^{1/2-\varepsilon}\right] &\leq (2n^{1/2-\varepsilon}+1)^b \max_{\vec{t}} \mathbb{P}\left[\vec{T}^{(2)} = \vec{t}\right] \\ &= \theta(n^{b(1/2-\varepsilon)})O(n^{-b(1-\beta)/2}) \\ &= O(n^{-b(\varepsilon-\beta)/2}) \\ &= o(n^{-a}).\end{aligned}$$

□

4.3.4 Proof of asymmetry regime

In the previous subsections we presented all the pieces needed to carry out the proof of our main result of this section, Theorem 4.8.

Proof of Theorem 4.8. Let $G = (V, E)$ be a $G(n, p)$ random graph. By the union bound, it is enough to prove that any two distinct vertices v_1, v_2 are locally symmetric with probability $o(n^{-2})$.

Let $v_1, v_2 \in V$ be arbitrary distinct vertices, and denote by X_1 the set of neighbors of v_1 that are not neighbors of v_2, by X_2 the set of neighbors of v_2 that are not neighbors of v_1, and by Y the set of common neighbors of v_1 and v_2. Additionally, denote by C_1 the number of edges between X_1 and Y, and by C_2 the number of edges between X_2 and Y. Our goal is to show that, with probability $o(n^{-2})$, $\mathcal{N}[v_1] = G[X_1 \cup Y]$ and $\mathcal{N}[v_2] = G[X_2 \cup Y]$ are not isomorphic. Note that these two graphs are not independent, since they share $G[Y]$ as a subgraph. We will show that, even if $G[Y]$ is removed from these two subgraphs, their degree sequences have large enough edit distance, by Theorem 4.24, that the effect of reinserting $G[Y]$, which is bounded by Corollary 4.16, is not enough to make these degree sequences equal, which implies $\mathcal{N}[v_1]$ and $\mathcal{N}[v_2]$ cannot possibly be isomorphic.

Before proceeding to this, we will need a few concentration bounds that will help us carry out our proof. First, note that $v_2 \in X_1$ and $v_1 \in X_2$ if and only if v_1 and v_2 are neighbors, and that all other vertices belong to X_1, X_2 and Y independently from each other, with probabilities $p(1-p)$, $p(1-p)$ and p^2, respectively. Thus, $|X_1|, |X_2| \leq_d \mathrm{Bin}(n-2, p) + 1$ and $|Y| \overset{d}{\sim} \mathrm{Bin}(n-2, p^2)$. By a similar reasoning, given X_1, X_2, and Y, it holds that $C_1 \overset{d}{\sim} \mathrm{Bin}(|X_1||Y|, p)$ and $C_2 \overset{d}{\sim} \mathrm{Bin}(|X_2||Y|, p)$.

Now, set $0 < \varepsilon_1 < 1$ constant, $\varepsilon_2 = \log n$, and $\varepsilon_3 > 0$ constant. Define the

following events:

$$A_1 = \{(n-2)p(1-p)(1-\varepsilon_1) < |X_1|, |X_2| < (n-2)p(1-p)(1+\varepsilon_1)\},$$
$$A_2 = \{|Y| < (n-2)p^2(1+\varepsilon_2)\},$$
$$A_3 = \{C_1, C_2 < sp(1+\varepsilon_3)\},$$

where $s = (n-2)^2p^3(1-p)(1+\varepsilon_1)(1+\varepsilon_2)$.

We will show that $\mathbb{P}(A_1, A_2, A_3) \geq 1 - o(n^{-2})$. First, by the Chernoff bound,

$$\begin{aligned}\mathbb{P}(\overline{A_1}) &\leq 2\exp\{-(n-2)p(1-p)\varepsilon_1 \min(1,\varepsilon_1)/3\} \\ &\quad + 2\exp\{-(n-2)p(1-p)\varepsilon_1 \min(1,\varepsilon_1)/2\}, \\ \mathbb{P}(\overline{A_2}) &\leq \exp\{-(n-2)p^2\varepsilon_2 \min(1,\varepsilon_2)/3\}.\end{aligned}$$

Furthermore, in the event $A_1 \cap A_2$, it holds that $|X_1||Y|, |X_2||Y| \leq s$, which implies $C_1, C_2 \leq_d \mathrm{Bin}(s,p)$ and, again by the Chernoff bound:

$$\mathbb{P}(A_1, A_2, \overline{A_3}) \leq 2\exp\{-sp\varepsilon_3 \min(1,\varepsilon_3)/3\}.$$

Note that the upper bounds for $\mathbb{P}(\overline{A_1})$, $\mathbb{P}(\overline{A_2})$ and $\mathbb{P}(A_1, A_2, \overline{A_3})$ are all $o(n^{-2})$, since $(n-2)p \geq \omega(\log n)$ and $(n-2)p^2\varepsilon_2 \geq \omega(1)$ by hypothesis, and $sp \geq \omega(\log n)$ as a consequence. Thus, $\mathbb{P}(A_1, A_2, A_3) \geq 1 - o(n^{-2})$.

We can now resume the main thread in our proof and show that, conditional on $A_1 \cap A_2 \cap A_3$, v_1 and v_2 are locally asymmetric with probability $1 - o(n^{-2})$. Recall that v_1 and v_2 are locally symmetric only if $\Delta(\mathcal{N}[v_1], \mathcal{N}[v_2]) = 0$. We can assume that v_1 and v_2 have the same degree — i.e. $|\mathcal{N}(v_1)| = |\mathcal{N}(v_2)|$, otherwise $\mathcal{N}[v_1]$ and $\mathcal{N}[v_2]$ have vertex sets of different sizes, which implies v_1 and v_2 are locally asymmetric. Note that this implies that $|X_1| = |X_2|$: if v_1 and v_2 are adjacent, then $\mathcal{N}(v_1) = X_1 \uplus Y \uplus \{v_2\}$ and $\mathcal{N}(v_2) = X_2 \uplus Y \uplus \{v_1\}$; if they are not adjacent, then $\mathcal{N}(v_1) = X_1 \uplus Y$ and $\mathcal{N}(v_2) = X_2 \uplus Y$.

Let $N_1 = (V_1, E_1) = G[\mathcal{N}(v_1) \setminus \{v_1\}]$ and $N_2 = (V_2, E_2) = G[\mathcal{N}(v_2) \setminus \{v_2\}]$, i.e., $\mathcal{N}[v_1]$ and $\mathcal{N}[v_2]$ with their centers removed. Since v_1 and v_2 are universal vertices of their respective neighborhoods, their removal keeps the edit distance between them unchanged, that is,

$$\Delta(\mathcal{N}[v_1], \mathcal{N}[v_2]) = \Delta(N_1, N_2).$$

Since Δ is a semimetric, it further holds that

$$\Delta(N_1, N_2) \geq \Delta(G[X_1], G[X_2]) - \Delta(G[X_1], N_1) - \Delta(G[X_2], N_2).$$

To lower bound the first term on the right-hand side, we note that the existence of edges between vertices in X_1 happens with probability p independently of the particular vertex pair, thus $G[X_1]$ is a $G(|X_1|, p)$ random graph. Since $p = \omega(n^{-1/2+\delta_1})$, by hypothesis, and $|X_1| = \theta(np)$ in $A_1 \cap A_2 \cap A_3$, it holds, for n large enough and for constants $c, c' > 0$, that

$$\begin{aligned}\frac{p}{\log |X_1|/|X_1|} &\geq \frac{p \cdot cnp}{\log c'np} \\ &\geq \frac{\omega(n^{2\delta_1})}{\log c'n} \\ &\geq \omega(1).\end{aligned}$$

It also holds that $|X_1|p^2 = p \cdot \theta(np^3) = o(1)$. Together, these inequalities imply $\omega(\log |X_1|/|X_1|) \leq p \leq o(|X_1|^{-1/2})$, and therefore $G[X_1]$ satisfies the hypotheses of Theorem 4.24. An analogous argument implies that $G[X_2]$ also satisfies these hypotheses.

Furthermore, note that $G[X_1]$ and $G[X_2]$ are independent, since their vertex sets are disjoint. This allows us to apply the results of Theorem 4.24, picking fixed $a = \frac{14}{1-14\delta_2}$ and $\varepsilon < \frac{49\delta_2}{8-14\delta_2}$:

$$\begin{aligned}\mathbb{P}[\Delta(G[X_1], G[X_2]) \geq |X_1|^{1/2-\varepsilon} | A_1 \cap A_2 \cap A_3] &= 1 - o((|X_1|p)^{-a}) \\ &= 1 - o((np^2)^{-a}) \\ &\geq 1 - o(n^{-a(1+2(-3/7-\delta_2))}) \\ &= 1 - o(n^{-a(1/7-2\delta_2)}) \\ &= 1 - o(n^{-2}).\end{aligned}$$

Note that these choices of a and ε are always possible and valid, since the hypotheses imply $\delta_2 < 1/14$, which makes $\frac{49\delta_2}{8-14\delta_2} \geq 7\delta_2 > 0$ and $\frac{14}{4-7\delta_2} > 0$. The two bounds on $|X_1|$ from the definition of event A_1 imply that, with probability $1 - o(n^{-2})$,

$$\Delta(G[X_1], G[X_2]) \geq \Omega((np)^{1/2-\varepsilon}).$$

To upper bound the remaining two terms, we begin by applying Corollary 4.16:

$$\Delta(G[X_1], N_1) \leq |V_1 \setminus X_1| + |C(X_1)|.$$

Now, all vertices of Y are in $V_1 \setminus X_1$, and so is v_2 if it is a neighbor of v_1 in G. All other vertices of G either are in X_1 or are not in V_1, thus $|V_1 \setminus X_1| \leq |Y| + 1$. Furthermore, $|C(X_1)|$ counts all edges from X_1 to $V_1 \setminus X_1$, regardless of whether v_2 belongs to $V_1 \setminus X_1$, since no edge connects X_1 to v_2 by construction. Thus, $|C(X_1)|$

counts edges from X_1 to Y, which implies $|C(X_1)| = C_1$. It follows that

$$\begin{aligned}\Delta(G[X_1], N_1) &\leq |Y| + 1 + C_1 \\ &\leq (n-2)p^2(1+\varepsilon_2) + (n-2)^2p^4(1-p)(1+\varepsilon_1)(1+\varepsilon_2)(1+\varepsilon_3) \\ &= \theta(n^2p^4 \log n).\end{aligned}$$

By an analogous argument, the same inequality holds for $\Delta(G[X_2], N_2)$. Now, note that

$$\begin{aligned}\frac{\Delta(G[X_1], N_1) + \Delta(G[X_2], N_2)}{\Delta(G[X_1], G[X_2])} &\leq \frac{\theta(n^2p^4 \log n)}{\Omega((np)^{1/2-\varepsilon})} \\ &= o(n^{3/2+\varepsilon}p^{7/2+\varepsilon} \log n) \\ &\leq o(n^{3/2+\varepsilon}n^{(-3/7-\delta_2)(7/2+\varepsilon)} \log n) \\ &= o(n^{(4-7\delta_2)\varepsilon/7-7\delta_2/2} \log n)\end{aligned}$$

The exponent of n in this expression satisfies

$$\begin{aligned}\frac{(4-7\delta_2)\varepsilon}{7} - \frac{7\delta_2}{2} &< \frac{4-7\delta_2}{7} \cdot \frac{49\delta_2}{8-14\delta_2} - \frac{7\delta_2}{2} \\ &= \frac{7\delta_2}{2} - \frac{7\delta_2}{2} = 0,\end{aligned}$$

thus the left-hand side is $o(1)$. As a consequence, $\Delta(\mathcal{N}[v_1], \mathcal{N}[v_2]) = \Delta(N_1, N_2) \geq \Delta(G[X_1], G[X_2])(1 + o(1)) \geq \omega(1)$. Therefore, under the condition $A_1 \cap A_2 \cap A_3$, with probability $1 - o(n^{-2})$, v_1 and v_2 are locally asymmetric. □

Figure 4.2 compares the results we have proven in the previous sections about local symmetry to corresponding results known for global symmetry. We note that, although omitted in this figure, global symmetry eventually reemerges for large enough average degree (more specifically, when $n(1-p) = o(\log n)$). We conjecture that the regimes for local asymmetry will follow a similar behavior of two phase transitions. The first one will happen between the currently identified regimes for local symmetry and local asymmetry, that is, for some average degree between $n^{1/3}$ and $n^{1/2}$. The techniques we used for identifying local symmetry regimes can be further explored, which leads us to believe this phase transition is more likely to happen closer to $n^{1/2}$ than to $n^{1/3}$ (possibly even precisely at $n^{1/2}$). The second phase transition would coincide with the one for global symmetry, thus happening around average degree $n - \log n$. This would happen because, for random graphs with such large average degree, the diameter of the network and average distance

between vertices will be very small. This leads to most neighborhoods containing nearly all vertices of the graph, effectively making local symmetry equivalent to global symmetry.

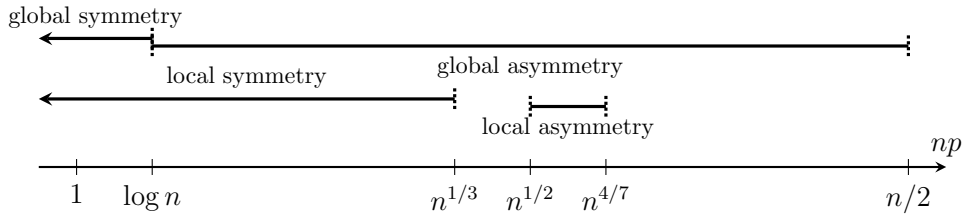


Figure 4.2: Comparison between local and global symmetry and asymmetry regimes in the $G(n, p)$ model.

4.4 Experimental results

We complement the previous theoretical results with an experimental evaluation on $G(n, p)$ random graphs. This evaluation consists of a numerical estimation, for a large range of values of n and p, of the probability of obtaining a locally symmetric $G(n, p)$, in contrast with a globally symmetric $G(n, p)$. Several libraries can be used for this purpose, among which we highlight Boost [40], which provides a simple routine to find an isomorphism between two graphs (if it exists), and the acclaimed Nauty [22], which not only allows for finding isomorphisms between two graphs but also for computing the automorphism group and the canonical form[4] of any graph, and is even recommended by Boost itself on grounds of efficiency.

However, since both libraries intend to determine automorphisms or isomorphisms with total effectiveness (that is, without a chance of undetection), this will result in a relatively high computational cost in the execution of these algorithms. This high cost for a single execution will scale with the graph sizes or number of executions we shall need:

- Determining global symmetry of a graph G requires the computation of a non-trivial automorphism in G, while determining local symmetry of G potentially requires performing an isomorphism test between neighborhoods for each pair of vertices in G;
- Since we intend to estimate the probability of having symmetry (local or global) in the $G(n, p)$ model, it is imperative that, for each pair of values

[4]The *canonical form* [41] of a graph G is a labelled graph Canon(G), isomorphic to G, such that two graphs G and H are isomorphic if and only if their canonical forms are precisely the same graph. Also known as *canonical labelling*.

for the n and p parameters, several graph instances are generated, enough to result in a relatively narrow confidence interval; in other words, we need to run the algorithms in the previous item multiple times for each pair of parameters.

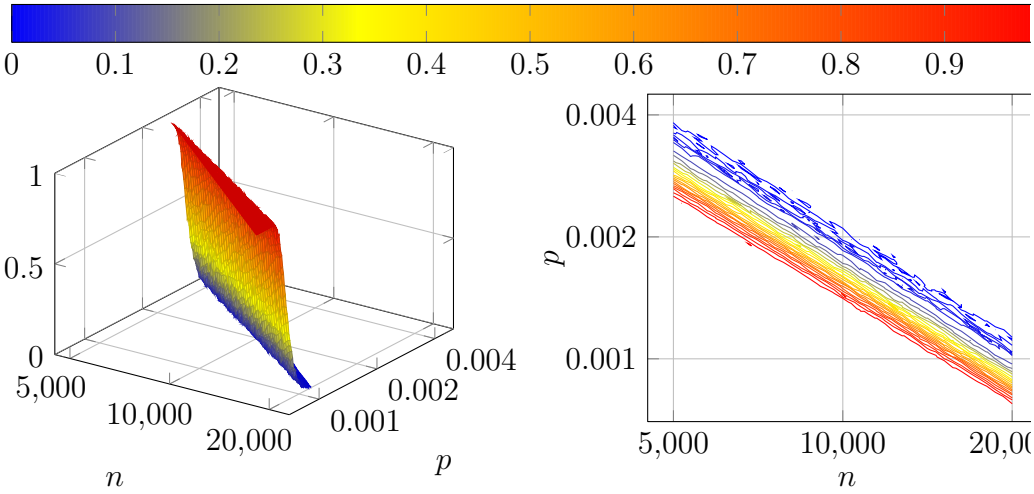
Our first approach has been to write simulations for the $G(n, p)$ model using the Nauty library. However, we observed that the practical impact of the issues previously described on the computational cost of our experimental setting have outweighted the benefit obtained with them and, for this reason, we have decided to devise an alternative experiment to be used to numerically approximate the desired probabilities.

In our approximation experiments, we will reduce our computational analysis of the graph structures to the comparison of specific degree sequences. More precisely, when determining whether two vertices u and v of a graph G are globally symmetric, we will check whether the degree sequence of neighbors of u is equal to the degree sequence of neighbors of v. For the purposes of this section, we call these sequences the *global degree sequences (GDS)* around u and v, respectively. Note that equality of global degree sequences is a necessary but not sufficient condition for global symmetry of vertices, thus resulting in an approximation from above of the true probabilities of symmetry, and the approximation takes place precisely by considering this condition as sufficient.

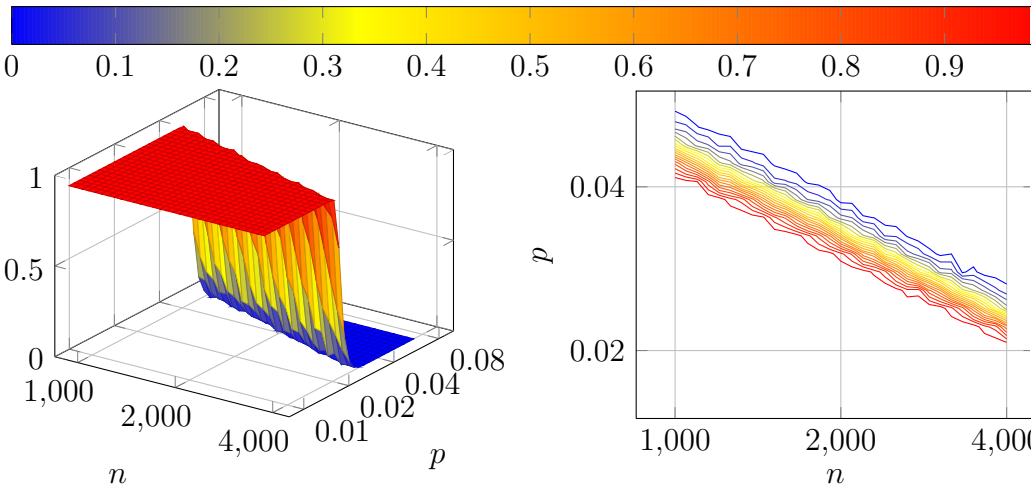
Similarly, when determining whether u and v are locally symmetric, we will check whether their the degree sequence of neighbors of u *in* $\mathcal{N}[u]$ is equal to the degree sequence of neighbors of v *in* $\mathcal{N}[v]$. Since we only consider edges inside the neighborhoods of u and v in these calculations, we call these the *local degree sequences (LDS)* around u and v, respectively. Once again, equality of local degree sequences is a necessary but not sufficient conditional for local symmetry of vertices, resulting in an approximation from above.

It is important to note that our findings from the previous sections make us believe that the approximation procedure is quite precise for local symmetry in those regimes previously analyzed: the proof of Theorem 4.5 (local symmetry regime) implies that a $G(n, p)$ random graph has fewer different LDSs than vertices, and the proof of Theorem 4.8 (local asymmetry regime) proceeds effectively shows that all vertices of a $G(n, p)$ random graph have different LDSs with probability at least $1 - o(1)$. Furthermore, verifying whether two vertices have the same global (or local) degree sequence has much lower computational cost, since it is only needed to order the degree sequences of both vertices and do a pairwise comparison of the elements of these sequences. It is still, however, a relatively high cost, which hinders our ability to perform these calculations on multiple instances of arbitrarily large graphs.

Considering the equality of these (global or local) degree sequences as an approximation to the occurence of (global or local) symmetry, and incorporating that to our vocabulary in the upcoming discussion, we can now proceed to the results of our experimental evaluation, as presented in the following figures:



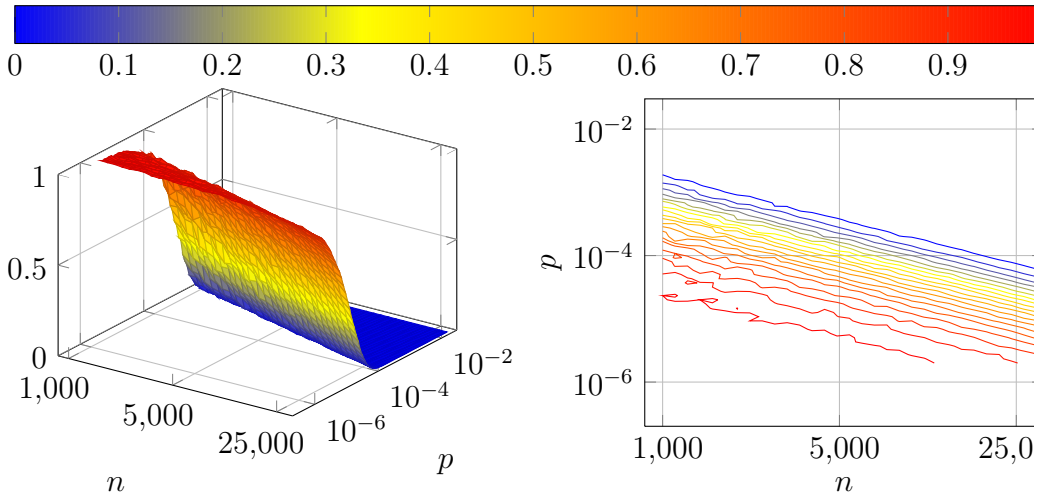
(a)



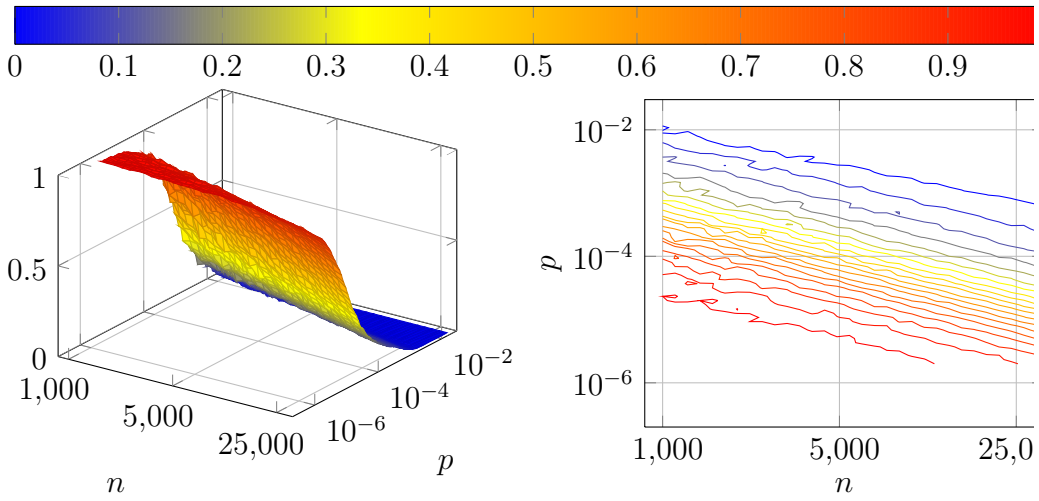
(b)

Figure 4.3: Estimated probabilities, as a function of the parameters n and p, of obtaining in a $G(n, p)$ at least one pair of vertices with the same GDS (Figure 4.3a) or with the same LDS (Figure 4.3b). On the graphs on the right-hand side, we highlight projections of the contour lines (for levels $0.05, 0.1, \ldots, 0.95$) on the corresponding surface on the left-hand side. Estimations obtained after 897 experiments in Figure 4.3a, and after 305 experiments in Figure 4.3b.

On Figure 4.3, we present the probabilities, experimentally estimated, of observing in a $G(n, p)$ graph at least one pair of vertices with the same global degree sequence (Figure 4.3a) or the same local degree sequence (Figure 4.3b). For convenience, the contour lines presented at the right-hand side are replicated in the first graph of Figure 4.5.



(a)



(b)

Figure 4.4: Estimated probabilities, as a function of the parameters n and p, of a pair of vertices, drawn uniformly at random from a $G(n, p)$, having the same GDS (Figure 4.4a) or the same LDS (Figure 4.4b). On the graphs on the right-hand side, we highlight projections of the contour lines (for levels $0.05, 0.1, \ldots, 0.95$) on the corresponding surface on the left-hand side. Estimations obtained after 50 experiments, with $\lfloor \frac{n}{100} \rfloor$ pairs of vertices extracted from each experiment (total of 500–12500 samples for each pair of parameters).

The first fact that we can notice is that there seems to have a relatively abrupt transition between the symmetry regimes (red contour lines) and the asymmetry regimes (blue contour lines), for both kinds of symmetry. Despite the relatively low number of vertices in the generated graphs (between 1000 and 20000), this phenomenon indicates the existence of a phase transition in the emergence of local asymmetry, similar to that of local symmetry.

We also notice the existence of a non-negligible gap between the contour lines corresponding to global symmetry and the ones corresponding to local symmetry.

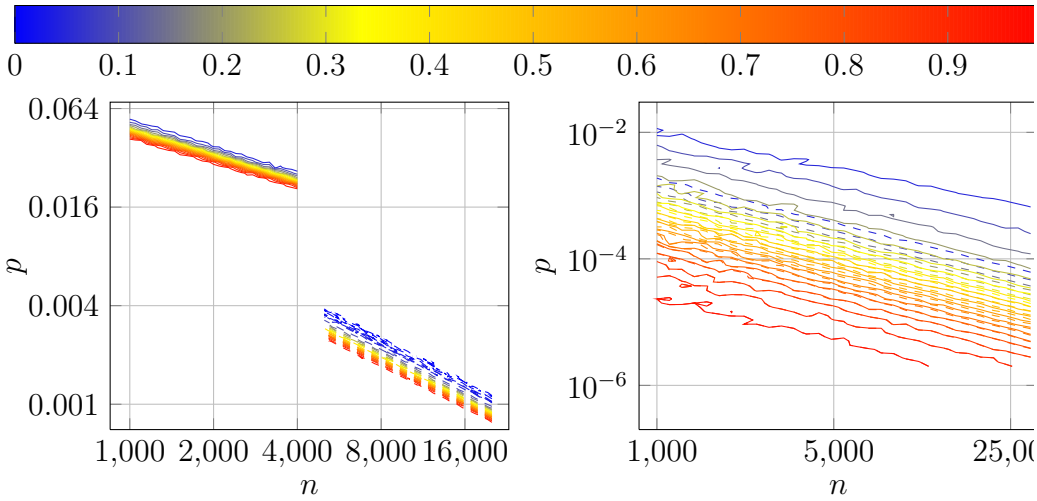


Figure 4.5: Superposition of contour lines presented in Figure 4.3 and Figure 4.4. Dashed curves correspond to calculations for same GDS, and continuous curves, for same LDS. The color coding matches previous figures.

In particular, we note that the highlighted contour lines for local symmetry occur for values of p about 5 times larger than the contour lines for global symmetry. This corroborates our theoretical findings that asserted the existence of intermediate density regimes in which we have, simultaneously, local symmetry and global asymmetry (region between the two sets of contour lines).

For the purposes of comparison, we performed a second set of experiments. In these experiments, instead of estimating the probability of having, in a $G(n, p)$, a pair of vertices with the same (global or local) degree sequence, we'd like to estimate these probabilities for a pair of vertices in a $G(n, p)$, selected uniformly at random. The results are presented in Figure 4.4, with contour lines replicated in the second graph of Figure 4.5.

This time, we notice a smoother transition between symmetry and asymmetry regimes, both in the global and local cases. This results in the emergence of global and local asymmetry occurring in a smoother manner, without the clear identification of a mixed regime.

Chapter 5

Network matching

"Tear down all my veils
Tear down 'til you see my naked core"
(Matias Kupiainen)

5.1 Literature review

Unlike the problem of symmetry in networks, which is a naturally analytical question to pose, the problem of *network matching* (or *graph matching*) is inspired by its practical applications. It has been described by several names, including graph reconciliation, approximate graph matching, graph alignment and inexact graph matching. In its most abstract formulation, one is given a pair of graphs, say $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$, and a hidden correspondence (or *matching*) between their vertex sets V_1 and V_2. The ultimate goal is to unveil this matching, thus identifying vertices of one graph in the other.

This problem relates to the previously mentioned *graph isomorphism* problem, which also takes two graphs G_1 and G_2 as input and consists of deciding whether there is an isomorphism between G_1 and G_2. However, while the graph isomorphism problem has a established formulation and approaches the idea of comparing two graphs' structures from a mostly theoretical point of view, the network matching problem has a more flexible formulation and is concerned about practical performance in real networks. The common thread in most of the work in network matching is the assumption that there exists a matching (either full or partial) between the vertex sets of both networks, meaning that the identities of corresponding vertices in opposite networks are the same, even though their labels are not. This induces a correlation between the structures of the two networks, which can be leveraged to uncover this underlying matching.

From this point of view, one can think of the identity correspondence as a feature of the pair of networks that is not perceivable or measurable directly, yet induces a

structural correlation which can, in turn, be perceived and measured. Note how this role is readily reversed with respect to the graph isomorphism problem, in which the existence and detection of a structural correlation — in this case, total correlation given by an isomorphism — induces a correspondence of identities, interpreted as an equal structural positioning akin to global symmetry. In other words, while in the graph isomorphism problem, structural correspondence *defines* a correct answer, in the network matching problem, it *evinces* a correct answer.

This distinction can be further noticed in that, in the network matching problem, we seek a single, specific correct matching, which implies that any alternative matchings with a consistent structural correspondence make the correct one harder to find. In contrast, in the graph isomorphism problem, such alternative matchings make the problem easier, since each and every correspondence that induce a perfect structural correlation is an equally correct answer.

A number of different applications can be mapped to the graph matching, the most recent and visible one being the study of privacy breach via unintended information leakage [42]. A standard scenario considers two social networks — e.g. the online social networks Facebook, Google+ and LinkedIn, — on which people fill in their profiles build relationships, such as friendship on Facebook or connection on LinkedIn. The structure of these networks is a rich object studied from several perspectives within the field of network science. This makes it natural to desire that data about these networks becomes available.

However, these networks comprise extremely sensitive and personal data, from simple metadata such as e-mail addresses and credit card numbers to more complex data such as mobility patterns and search history, which is not supposed to be made available to the public but whose study is still of value and interest. The most usual precaution to protect privacy is to erase all personal data prior to the release of the networks at hand, thus protecting the identity of their nodes. In this context, matching the structures of both networks allows one to cross-reference their metadata, which constitutes both a powerful source and an undesired leakage of personal information.

In other contexts, matching two networks is not a dangerous procedure to be prevented, but a technique to obtain previously unknown or undetected information. This is the case of bioinformatics [43], where matching protein-protein interaction networks (PPI networks) can lead to the identification of the same protein, or functionally similar proteins, in different species, regardless of how they have been originally labelled. This can shed light on the protein evolutionary pattern across generations of related species. Similar procedures can also be applied to gene-regulatory, metabolic and other biological networks [44]. Other examples include computer vision [28], where network representations of images can be matched for the identi-

fication of equivalent physical features [45], and ontology networks [46, 47], where two knowledge bases can be combined to allow for their interoperability, build more complete knowledge bases or handle ambiguity issues.

Most work in network matching has focused on the design of algorithms to perform the matching of real networks [48], sometimes using just the structure of the networks themselves, sometimes using node features as additional information sources [49]. Such features could be, for instance, sections of a protein's amino acid sequence, or the lexical content of entries in a knowledge base. In most real networks, corresponding vertices are more likely to have similar features, which allows algorithms to leverage this information in order to build a more accurate matching.

On the importance of structures, a landmark work is Narayanan and Shmatikov's pairwise matching of online social networks such as Twitter, Flickr and LiveJournal [50], believed to be the first successful large-scale network de-anonymization attack based solely on the networks' structure. Two main elements drive their technique: the identification of a small number of trusted, correctly matched pairs called *seeds*, followed by an iterative process of *propagation* which expands this small partial match into a full match of the two networks at hand.

Their instance of this framework performs both steps by using heuristics: seed pairs are found by searching for cliques of known fixed size and matching their vertices using their degrees and codegrees[1]; the propagation step looks at neighbors of previously matched vertices, assigns scores to potential new matches based on the fraction of matched neighbors and accepts matches with a high, heuristic strength factor, which takes into account the degrees of matched vertices, the eccentricity of the match's score, among other factors. The size of initial seed set is one of the most sensitive parameters with respect to the algorithm's performance.

This framework has spawned a number of similar variations [51, 52]. Kazemi et al. [53] expanded on this by incorporating the notion of *untrusted seeds*, which can be used to move the propagation process forward in cases where it would die out otherwise. Similar algorithms based on probabilistic approaches have also been proposed. Pedarsani et al. [54] take an iterative Bayesian approach, starting with matching nodes with highest degree and using half of the matches from previous iterations as evidence in maximizing posterior probabilities of matching at each step. Doshi and Thomas [47] follow a similar path, growing a matching from a set of seeds and modeling the propagation process as an expectation-maximization (EM) iterative procedure. However, the prior matching probabilities used in their algorithm take node attributes into account, which makes this algorithm not purely based on network structure.

In contrast, other works have focused on a more rigorous mathematical model-

[1]The *codegree* of vertices u and v is the number of common neighbors of u and v.

ing and analysis of the problem. This usually requires not only a precise formulation of the problem itself, but also a mathematical model for the input graphs to be matched that captures their structural similarity. For instance, Pedarsani and Grossglauser [27] introduced the $G(n, p, s)$ model to study fundamental achievability limits for network matching — we will detail this particular model in the next section.

5.2 The $G(n, p, s)$ random graph model

The $G(n, p, s)$ model has been introduced by Pedarsani and Grossglauser [27] as a model for structurally correlated random graphs. The idea behind this model is that two correlated networks, such as the Facebook and Twitter social networks can be seen as noisy observations of a true, underlying network — in this example, the real-world friendship social network. This is captured in a random graph framework by using an intermediate *generator graph* $G = (V, E)$, which is then subjected to independent applications of noise, in the form of edge insertion and removal, to obtain two *observed graphs*, $G_1 = (V, E_1)$ and $G_2 = (V, E_2)$. We can also say that G_1 and G_2 are independent random *samples* of G. This process is illustrated in Figure 5.1.

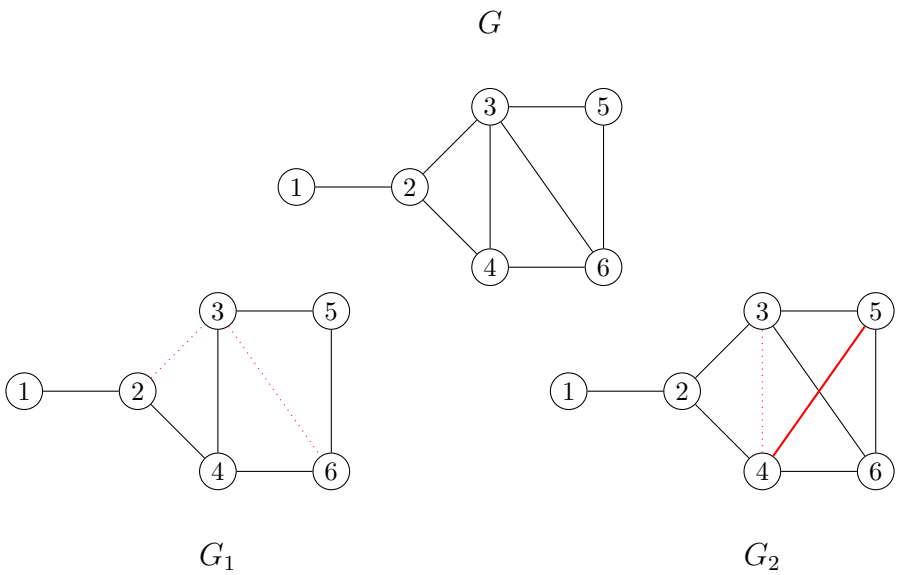


Figure 5.1: Example of graphs G_1 and G_2, correlated through a generator graph G. Structural changes due to noise are highlighted in red: thick edges have been inserted, dashed edges have been removed. Note that, since inserted noise is small, the structures of G_1 and G_2 are similar.

The network matching problem then reduces to unveiling the hidden mapping between G_1 and G_2 induced by the generator graph G, that is, to determine which

pairs of vertices, one in G_1 and one in G_2, are images of the same vertex of G. Note that, by construction of the model, this mapping trivially corresponds to the identity permutation over V (which we denote by $\mathbb{I}_V$), but such trivial determination of $\mathbb{I}_V$ requires access to the labels of each vertex in both G_1 and G_2. Therefore, what we seek is a joint structural statistic of G_1 and G_2 — i.e., some function of the structures of both G_1 and G_2, and of a permutation between their vertex sets — which is label-invariant and makes $\mathbb{I}_V$ stand out. This could be, for instance, a real-valued statistic which is maximized by $\mathbb{I}_V$ with high probability.

A variation of this framework incorporates label removal as an additional step performed after the sampling process, by applying a fixed permutation $\pi \in \mathrm{Sym}(V)$[2] to the vertex set of G_2 and applying the corresponding action to the edge set of G_2. In this variation, we are given the structures of G_1 and G_2 but not π, and our goal can be made even more precise: to estimate π based on the structures of G_1 and G_2. As we will show later, this allows for a formalization of this statistical inference problem, and for this reason, we will call this variation the *inference version* of the model, with the previous model, without the label removal step, being called the *standard version*. The standard version is more suitable for probability analysis, due to its simpler notation, and will therefore be the version we discuss throughout this thesis unless otherwise noted. In practical purposes, since the problem is only meaningful in the absence of vertex labels, both models are equivalent.

Now, the $G(n, p, s)$ model is a particular realization of this framework which takes G to be an Erdős-Rényi random graph with size n and edge probability p, and the sampling process to be an independent removal of the edges of E, where each edge is kept with probability s and removed with probability $1 - s$. The parameter s controls the structural similarity of the observed graphs G_1 and G_2. Note that both are $G(n, ps)$ random graphs, since the presence of an edge in an observed graph is independent of other edges, and each edge must "survive" two independent Bernoulli trials, with probabilities p (generation of G) and s (sampling process). However, the observed graphs are not independent (if $0 < p < 1$), but correlated — in particular, if we know a certain edge e belongs to E_1, e must belong to E, which implies e belongs to E_2 with probability s, rather than ps.

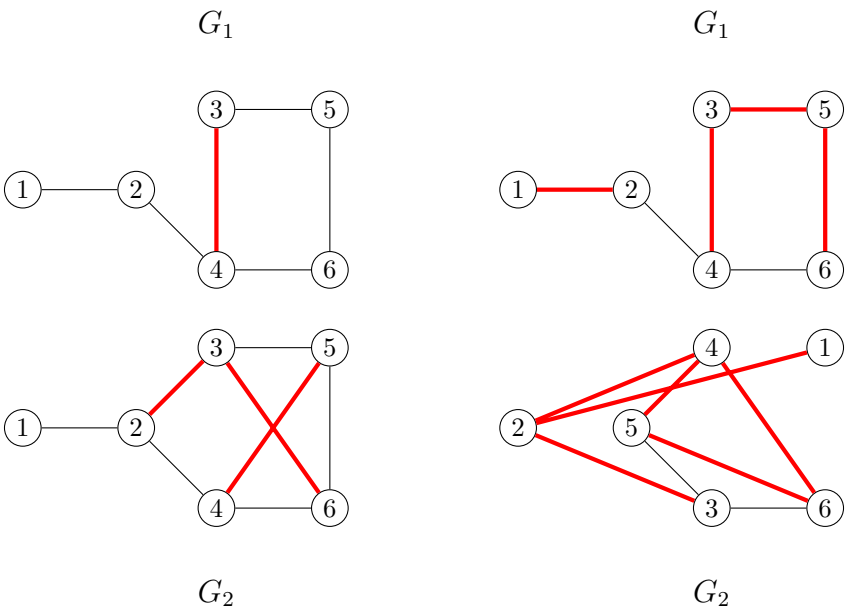
Recent work has shown some very interesting results for this model. For instance, define the *edge mismatch* function as follows:

$$\Delta(\pi) = \Delta(\pi; G_1, G_2) = \sum_{e \in \binom{V}{2}} \mathbb{1}_{\{e \in E_1 \otimes \pi(e) \in E_2\}},$$

where $\otimes$ denotes the binary exclusive-or operator. This function counts how many

[2] $\mathrm{Sym}(V)$ denotes the symmetric group on V.

edges are present in only one of G_1 and G_2, when mapping their vertex sets by π. An example of this calculation is given in Figure 5.2.



(a) Mapping G_1 into G_2 using identity permutation $\mathbb{I}_V$

(b) Mapping G_1 into G_2 using permutation $\pi = (1\ 2\ 5)(3\ 4)(6)$

Figure 5.2: Calculating edge mismatch between graphs G_1 and G_2, presented in previous example. Mismatched edges have been highlighted. π is given in cycle notation. In this example, $\Delta(\mathbb{I}_V) = 4$ and $\Delta(\pi) = 10$.

Intuitively, one would expect this edge mismatch to be small for $\mathbb{I}_V$ and similar permutations, since correct mapping of vertices induce correct mapping of vertex pairs, in which presence of edges is correlated. Indeed, it has been shown that, if $nps^2 = 2\log n + \omega(1)$ (and some additional conditions), then $\mathbb{I}_V$ uniquely minimizes $\Delta(\pi)$ a.a.s. [55, Theorem 2] Even more, consider the inference version of $G(n, p, s)$ and denote by $\pi_{MLE} = \pi_{MLE}(G_1, G_2)$ the maximum likelihood estimator for π. Then:

$$\pi_{MLE} = \arg\min_{\pi} \Delta(\pi; G_1, G_2).$$

This means that, not only the MLE for π is a.a.s. a consistent estimator if the expected degrees of G_1 and G_2 are large enough, but it can be inferred from a relatively straightforward statistic of the input random graphs. Of course, this says nothing about how to efficiently compute this statistic, but it implies that the network matching problem is feasible, given enough computational power.

On the other hand, if $nps = o(\log n)$, then there is no statistic $f(\pi; G_1, G_2)$ which can be used to distinguish $\mathbb{I}_V$ from the remaining permutations a.a.s. This comes from the fact that, in such sparse regime, $\mathrm{Aut}(G_1)$ and $\mathrm{Aut}(G_2)$ are a.a.s. non-trivial, which means any statistic f will satisfy $f(\mathbb{I}_V) = f(a_2 \circ \mathbb{I}_V \circ a_1)$, for any

$a_1 \in \mathrm{Aut}(G_1), a_2 \in \mathrm{Aut}(G_2)$. Cullina and Kiyavash [55, Theorem 2] showed that the same conclusion holds if $nps^2 = \log n - \omega(1)$.

Other models related to this framework have been investigated as well. For instance, Chiasserini et al. [56] have instantiated this framework with a similar edge-independent sampling process but with G being a power-law random graph (PLRG) rather than an Erdős-Rényi random graph. A slightly more generalized version of this had previously been used by Korula and Lattanzi [52] to study the peformance of a matching algorithm based on degree sequences. Cullina and Kiyavash [55] considered a generalization of $G(n, p, s)$ without an explicit generator graph, where an edge is present in G_1 and G_2 with possibly inhomogeneous probabilities given by $p_{ij} = \mathbb{P}[\mathbb{1}_{e \in E_1} = i, \mathbb{1}_{e \in E_2} = j | e \in E]$. Kazemi et al. [57] considered an extended framework in which vertices of G are also sampled to form the vertex sets of G_1 and G_2, thereby accounting for partial vertex overlap.

Chapter 6

Multiple network matching

"Out of the shadows
One by one they came
To shed their light upon his moment of doubt and pain"
(John Petrucci)

In this thesis, we will be interested in exploring the *multiple network matching* (or *multiple graph matching*) problem. It is a natural extension of the network matching problem for an input that comprises an arbitrary number of correlated graphs over a common vertex set with the goal of matching all graphs simultaneously. Work on this problem has followed a similar path as the network matching problem and focused on the development of heuristic algorithms, usually targeting specific applications [58] such as detection of function preservation in protein-protein interaction [59, 60] or pattern recognition in computer vision [61, 62].

To the best of our knowledge, no work has proposed to study the fundamental limits of this problem, in a fashion similar to traditional network matching on two graphs. There is an interesting intuitive trade-off taking place. On one hand, the larger number of graphs to be matched makes this problem more general, ergo more difficult to solve. However, these additional graphs also bring in additional information about the correlation between their structure, which can be leveraged in solving for the correct matching. Ultimately, we are interested in the following question: does the additional information obtained by having a larger number of network samples make the multiple network matching problem easier to solve, and if so, under which conditions?

6.1 Hyperpermutations

Before investigating this model further, we must devise a mathematical object to represent the result of the multiple matching problem. In the matching problem

with two graphs, it was enough to consider the permutations of the vertex set, as this set represents all possible mappings from one graph to the other graph. For multiple graph matching, we could simply consider an indexed sequence of pairwise permutations of the node set, corresponding to each pairwise mapping between our k graphs. However, these mappings are expected to be consistent. We choose to account for such consistency by design, through the following definition:

Definition 6.1. *A* hyperpermutation of order k *or* k-hyperpermutation Π *over a set X is a pairwise indexed family of permutations $\{\Pi_{ab}\}_{a,b\in[k]}$, $\Pi_{ab} \in \mathrm{Sym}(X)$, such that:*

$$\Pi_{bc} \circ \Pi_{ab} = \Pi_{ac} \qquad \forall\ a, b, c \in [k].$$

The set of all k-hyperpermutations over X will be called the *k-hypersymmetric group* and denoted by $\mathrm{Sym}_k(X)$. Whenever sufficiently explicit, the order k will be omitted from our verbal discussions. We can think of a k-hyperpermutation Π as if we had k copies of the set X and Π_{ij} represents the mapping between nodes on the i-th and the j-th copy. Each mapping Π_{ij} will be called a *marginal permutation* of Π. The most direct use of this object in the multiple network matching problem lets $X = V$ the common vertex set of all input graphs, in which case each hyperpermutation is a different way of collectively matching their vertex sets.

The consistency restrictions in Definition 6.1 ensure that, for any sequence of mappings between these copies, their compositions result on the same end-to-end mapping:

Remark 6.2. *For every $a, b \in [k]$, $\Pi_{aa} = \mathbb{I}_X$ and $\Pi_{ab} = \Pi_{ba}^{-1}$. Also, for every finite sequence $a_1, \cdots, a_t$, with $t \geq 3$:*

$$\Pi_{a_{t-1}a_t} \circ \cdots \circ \Pi_{a_2a_3} \circ \Pi_{a_1a_2} = \Pi_{a_1a_t}$$

Intuitively, this means that a hyperpermutation Π is uniquely defined by a fraction of its marginal permutations Π_{ij}, as long as they are carefully selected to provide a single chain between indices in $[k]$. If X is finite, elementary graph theory implies $k - 1$ such elements are necessary and sufficient. To see that, recall that a tree is a graph that is both connected and minimal about this property (removing any edges from it makes the resulting graph not connected).

Statement 6.3. *Let T be a tree with node set $V(T) = [k]$, and let $E(T)$ be its edge set. Then there is a bijection between $\mathrm{Sym}(X)^{E(T)}$ and $\mathrm{Sym}_k(X)$.*

Proof. We'll use a number of widely known properties of trees in this proof. For shortness, we'll simply state them as they are required.

Let $P = \{P_e\}_{e \in E(T)}$ be an arbitrary element of $\mathrm{Sym}(X)^{E(T)}$ (note that e represents a pair of copies of X), and Π be an arbitrary hyperpermutation in $\mathrm{Sym}_k(X)$. We'll say that P and Π *match* if the following property holds:

$$P_e = \Pi_e \ \forall \ e \in E(T).$$

For any Π, it is easy to see that there is a unique P such that P and Π match (such P is obtained by mere enumeration of marginal permutations $\{\Pi_e\}_{e \in E(T)}$). We'll show that the converse is also true: for any P, there is a unique Π such that P and Π match — or equivalently, there is a Π such that P and Π match (existence), and for two hyperpermutations Π, Π', if P and Π match and P and Π' match, then $\Pi = \Pi'$ (uniqueness). This implies the desired result, as the function $f : \mathrm{Sym}(X)^{E(T)} \to \mathrm{Sym}_k(X)$, where $f(P)$ is the unique Π such that P and Π match, is the desired bijection.

We begin by showing uniqueness of P. Suppose P and Π match, and that P and Π' match. Then it holds that $P_e = \Pi_e = \Pi'_e$ for every $e \in E(T)$. Now, take some pair $e = vw \notin E(T)$. Since T is a tree, there is a sequence $v_0, \ldots, v_j \in [k]$ of distinct elements such that $v_0 = v$, $v_j = w$ and $v_{i-1}v_i \in E(T)$ for every $i \in [j]$. We can, then, write:

$$\begin{aligned}
\Pi_e = \Pi_{vw} &= \Pi_{v_0 v_j} \\
&= \Pi_{v_{j-1} v_j} \circ \cdots \circ \Pi_{v_0 v_1} \\
&= \Pi'_{v_{j-1} v_j} \circ \cdots \circ \Pi'_{v_0 v_1} \\
&= \Pi'_{v_0 v_j} = \Pi'_{vw} = \Pi'_e
\end{aligned}$$

Therefore, $\Pi_e = \Pi'_e$ for every $e \in \binom{V}{2}$, that is, $\Pi = \Pi'$.

To show existence, consider some pair $e = vw \in \binom{[k]}{2}$. Again because T is a tree, the aforementioned path $v = v_0, \ldots, v_j = w \in [k]$ from v to w is unique satisfying $v_{i-1}v_i \in E(T)$ for every $i \in [j]$ and having unique elements. Construct Π by taking $\Pi_e = P_{v_{j-1}v_j} \circ \cdots \circ P_{v_0 v_1}$. If $v = w$, since the composition on the right-hand side is empty, take $\Pi_e = \mathbb{I}_X$.

All that remains is to show that Π is indeed a hyperpermutation, that is, Π satisfies the consistency restrictions.Take $a, b, c \in [k]$. If $a = b = c$, then consistency holds trivially as both sides equal $\mathbb{I}_X$. If two of these indices are equal, say $a = b$, then $\Pi_{ab} = \mathbb{I}_X$, both sides equal Π_{bc} and the identity is again true (similarly for $a = c$ and $b = c$).

If $a \neq b \neq c \neq a$, then in T there are unique paths $a = v_0, \ldots, v_j = b$ and $b = w_0, \ldots, w_{j'} = c$. If these paths don't overlap except for in b, we conclude that the composition $v_0, \ldots, v_j = w_0, \ldots, w_{j'}$ is a path in T from a to c with distinct

elements and, therefore, the unique such path. This implies $\Pi_{bc} \circ \Pi_{ab} = P_{w_{j'-1}w_{j'}} \circ \cdots \circ P_{w_0w_1} \circ P_{v_{j-1}v_j} \circ \cdots \circ P_{v_0v_1} = \Pi_{ac}$.

On the other hand, if the paths overlap, such overlap must be a contiguous subpath incident to b (otherwise, we could use the non-overlapping sections of both paths to build a cycle in T, which would contradict T being a tree). That is, if m is the number of overlapping edges, then it is true that $v_{j-1} = w_1, \ldots, v_{j-m} = w_m$ with all other elements in both paths being distinct among themselves and from the elements in the overlap. This means the unique path in T from a to c is the composition of non-overlapping sections $v_0, \ldots, v_{j-m} = w_m, \ldots, w_{j'}$. In this case, we can write:

$$\begin{aligned}
\Pi_{bc} \circ \Pi_{ab} &= P_{w_{j'-1}w_{j'}} \circ \cdots \circ P_{w_0w_1} \circ P_{v_{j-1}v_j} \circ \cdots \circ P_{v_0v_1} \\
&= P_{w_{j'-1}w_{j'}} \circ \cdots \circ P_{w_mw_{m+1}} \circ P_{w_{m-1}w_m} \cdots \circ P_{w_0w_1} \circ \\
&\quad \circ P_{v_{j-1}v_j} \circ \cdots \circ P_{v_{j-m}v_{j-m+1}} \circ P_{v_{j-m-1}v_{j-m}} \circ \cdots \circ P_{v_0v_1} \\
&= P_{w_{j'-1}w_{j'}} \circ \cdots \circ P_{w_mw_{m+1}} \circ P_{w_{m-1}w_m} \cdots \circ P_{w_0w_1} \circ \\
&\quad \circ P_{w_1w_0} \circ \cdots \circ P_{w_mw_{m-1}} \circ P_{v_{j-m-1}v_{j-m}} \circ \cdots \circ P_{v_0v_1} \\
&= P_{w_{j'-1}w_{j'}} \circ \cdots \circ P_{w_mw_{m+1}} \circ (P_{w_{m-1}w_m} \cdots \circ P_{w_0w_1} \circ \\
&\quad \circ P^{-1}_{w_0w_1} \circ \cdots \circ P^{-1}_{w_{m-1}w_m}) \circ P_{v_{j-m-1}v_{j-m}} \circ \cdots \circ P_{v_0v_1} \\
&= P_{w_{j'-1}w_{j'}} \circ \cdots \circ P_{w_mw_{m+1}} \circ P_{v_{j-m-1}v_{j-m}} \circ \cdots \circ P_{v_0v_1} = \Pi_{ac}
\end{aligned}$$

In both cases, then, the consistency restrictions is satisfied. Since a, b, c were arbitrary, all consistency restrictions are satisfied, and Π is indeed a hyperpermutation. This concludes the proof. □

To ease notation, we may perform this construction with more convenient trees. For instance, we can take $T = P_k$, a path graph over $[k]$, or $T = K_{1,k-1}$, a star graph with $k-1$ endpoints. These constructions have a clear interpretation under graph matching: taking $T = P_k$ is equivalent to arranging our k graphs in a sequence, and matching the node set of each graph and the next one in the sequence; taking $T = K_{1,k-1}$ is equivalent to selecting one of our k graphs as a reference, and matching its node set to the node sets of every other graph.

Among all hyperpermutations, a few in particular will be called upon in this work and should be singled out at this early stage. First, the *identity hyperpermutation* of order k over X, which we denote by $\mathbb{I}^k_X$, is the only permutation in $\mathrm{Sym}_k(X)$ satisfying the following property:

$$\left(\mathbb{I}^k_X\right)_{ab} = \mathbb{I}_X \ \forall \ a, b \in [k].$$

It is an intuitive extension of $\mathbb{I}_X$ for the case of multiple networks and, naturally, it is equal to $\mathbb{I}_X$ when $k = 2$.

Furthermore, we define the class $\mathcal{T}$ of *hypertranspositions*, which comprises all hyperpermutations which simply transpose two elements of X between two subsets of "copies" of X. Formally, a hyperpermutation Π is a k-hypertransposition over X if there are two non-empty subsets $A, B \subset [k]$ with $B = [k] \setminus A$, and two distinct elements $x, y \in X$ such that:

$$\Pi_{ab} = \begin{cases} (x\ y), & \text{if } a \in A, b \in B \text{ or vice-versa} \\ \mathbb{I}_X, & \text{otherwise} \end{cases}$$

Again naturally, this class reduces to the class of traditional transpositions when $k = 2$.

The following additional definition should prove to be useful throughout this chapter.

Definition 6.4. *Let X be a set and $\Pi \in \mathrm{Sym}_k(X)$ be a k-hyperpermutation over X. For any element $x \in X$, the* horizontal mapping *of x by Π, denoted by $\mathrm{Hor}_\Pi(x)$, is the sequence of elements of X given by $\{\Pi_{1j}(x)\}_{j\in[k]}$. The j-th element of $\mathrm{Hor}_\Pi(x)$ will be denoted by $[\mathrm{Hor}_\Pi(x)]_j$.*

In other words, the horizontal mapping of an element x in a hyperpermutation Π consists of all elements y such that x in the first copy of X is mapped to y in the j-th copy of X[1].

Horizontal mappings provide us with an alternative way to view hyperpermutations. In particular, we can take a hyperpermutation Π and arrange in a $|X| \times k$ matrix, with the first column of the matrix simply displaying an enumeration of X and each row consisting of the horizontal mapping of its first element (hence its name, *horizontal* mapping). We call this the *matrix form* of a hyperpermutation. Its construction implies the j-th column of the matrix form of Π corresponds to the permutation Π_{1j}, thus Π is uniquely defined from its matrix form. See Figure 6.1 for an example.

It's important to notice that, so far, we have a purely combinatorial object as we have not specified what is the nature of the set X over which we will apply hyperpermutations. In particular, when using this object in the context of random graphs, beyond the direct choice of considering hyperpermutations over V, we also would like to consider hyperpermutations over the set of potential edges $\binom{V}{2}$. In

[1]Note that our use of the first copy as the base reference in this definition does not impose any particularities on that copy. Using graph j as the base reference would yield the same set of objects, with the caveat that the horizontal mapping of x by the current definition would become the horizontal mapping of $\Pi_{1j}(x)$ by the alternative one.

$$\Pi_{12} = (a\ c)(b\ d\ e)(f)$$
$$\Pi_{23} = (a\ e\ c\ f)(b)(d)$$
$$\Pi_{34} = (a\ d\ f)(b\ e\ c)$$

(a) Description of Π, with marginals given in cycle notation.

X_1	X_2	X_3	X_4
a	c	f	a
b	d	d	f
c	a	e	c
d	e	c	b
e	b	b	e
f	f	a	d

(b) Matrix form of Π.

Figure 6.1: Matrix form of a hyperpermutation Π of order 4 over a set $X = \{a, b, c, d, e, f\}$. The horizontal mapping of each element of X by Π is displayed in a row of the matrix form of Π — for instance, $\mathrm{Hor}_\Pi(a) = (a, c, f, a)$. The i-th column of this matrix form is equal to the marginal permutation Π_{1i} expressed in one-line notation.

fact, hyperpermutations on a set of vertices naturally induce hyperpermutations on the corresponding set of potential edges via the actions of their marginals on $\binom{V}{2}$:

Lemma 6.5. *Let Π be a hyperpermutation over V, and for each $a, b \in [k]$, let $\Phi_{ab} : \binom{V}{2} \to \binom{V}{2}$ be the function such that $\Phi_{ab}(\{u, v\}) = \{\Pi_{ab}(u), \Pi_{ab}(v)\}$. Then, $\Phi = \{\Phi_{ab}\}_{a,b\in[k]}$ is a hyperpermutation over $\binom{V}{2}$.*

Proof. Each $\Phi_{a,b}$ is trivially a permutation over $\binom{V}{2}$, so it only remains to prove that Φ satisfies the consistency restrictions. Let $a, b, c \in [k]$ and $u, v \in V$. Since Π is a hyperpermutation, and therefore satisfies its own consistency restrictions, it holds that $\Phi_{bc} \circ \Phi_{ab}(\{u, v\}) = \Phi_{bc} \circ \{\Pi_{ab}(u), \Pi_{ab}(v)\} = \{\Pi_{bc} \circ \Pi_{ab}(u), \Pi_{bc} \circ \Pi_{ab}(v)\} = \{\Pi_{ac}(u), \Pi_{ac}(v)\} = \Phi_{ac}(u)$. As a, b, c, u, v were arbitrary, this proves the result. □

By a slight abuse of notation, whenever Π is a hyperpermutation over V, we will denote the hyperpermutation Φ on $\binom{V}{2}$, defined in the previous lemma, also by Π.

6.2 The $G_k(n, p, s)$ random graph model

Recall from our discussion in the beginning of this chapter that we would like to understand whether the additional information from multiple networks can make the problem of matching easier to solve. In order to start shedding light on this question, we propose a natural extension of the $G(n, p, s)$ model to account for the occurrence of any number of correlated graphs.

In this model, which we call the $G_k(n, p, s)$ model, k correlated random observed graphs $G_1 = (V, E_1), \ldots, G_k = (V, E_2)$, are generated by independently applying noise to a generator graph $G = (V, E)$, which is itself distributed as an Erdős-Rényi random graph. Noise again consists on the sampling of each edge in E with probability s, independently of other edges. Note that the $G(n, p, s)$ is a particular

case of this model for $k = 2$. We will use the following binary random variables to denote the most elementary events in this generation process:

- G_e denotes, for a pair of vertices $e \in \binom{V}{2}$, whether there is an edge between them in the generator G. By construction, $G_e \overset{d}{\sim} \mathrm{Be}(p)$;
- S_e^j denotes, for a pair of vertices $e \in \binom{V}{2}$ and a graph G_j with $j \in [k]$, whether the edge e, if present in G, is also present in G_j. Again by construction, $S_e^j \overset{d}{\sim} \mathrm{Be}(s)$.

All of these random variables are mutually independent. In particular, S_e^j is independent of the existence of e in G, but only influences the structure of G_j if e is present on the generator. Using the additional random variable $G_e^j = G_e S_e^j$, we can construct the edge sets of our observed graphs as:

$$\begin{aligned} E &= \left\{ e \in \binom{V}{2} \ : \ G_e = 1 \right\} \\ E_j &= \{ e \in E \ : \ S_e^j = 1 \} \\ &= \left\{ e \in \binom{V}{2} \ : \ G_e^j = 1 \right\} \end{aligned}$$

As in the case on only two graphs, the identity hyperpermutation $\mathbb{I}_V^k$ is the sought correct mapping of the vertex sets of all graphs. Therefore, what we seek is some structural statistic $f(\Pi)$ over all possible matchings that allows us to distinguish $\mathbb{I}_V^k$ from the remaining hyperpermutations. Generalizations of the mismatch $\Delta(\pi; G_1, G_2)$, which we have previously presented, are naturally strong candidates, and among these, it is most natural to formulate such statistics by counting mismatches in different ways. For instance:

- one can count edge mismatches pairwise between all graph pairs — we call this the *full edge mismatch* of Π:

$$\Delta_F(\Pi; G_1, \ldots, G_k) = \sum_{i,j \in [k]} \Delta(\Pi_{ij}; G_i, G_j) = \sum_{i,j \in [k]} \sum_{e \in \binom{V}{2}} \mathbb{1}_{\{e \in E_i \otimes \Pi_{ij}(e) \in E_j\}};$$

- one can fix one of the graphs, say G_1, and count edge mismatches between this and all other graphs — we call this the *star edge mismatch* of Π:

$$\Delta_S(\Pi; G_1, \ldots, G_k) = \sum_{j \in [k]} \Delta(\Pi_{1j}; G_1, G_j) = \sum_{j \in [k]} \sum_{e \in \binom{V}{2}} \mathbb{1}_{\{e \in E_1 \otimes \Pi_{1j}(e) \in E_j\}};$$

- one can take a tree $T = T(k)$ over $[k]$, as in Statement 6.3, and exploit its

structure by counting edge mismatches only over graph pairs that are represented by the tree — we call this the *T-edge mismatch* of Π:

$$\Delta_T(\Pi; G_1, \ldots, G_k) = \sum_{(i,j)\in E(T)} \Delta(\Pi_{ij}; G_i, G_j) = \sum_{(i,j)\in E(T)} \sum_{e\in\binom{V}{2}} \mathbb{1}_{\{e\in E_i \otimes \Pi_{ij}(e)\in E_j\}}.$$

Our ultimate goal is to use label-invariant structural statistics, such as these, to identify the correct matching, $\mathbb{I}_V^k$. While this leaves enough room for several precise statements that realize such identification, such statements are most naturally formulated as minimization or maximization problems — that is, one seeks to determine under which conditions a matching-dependent statistic such as Δ_T is minimized by $\mathbb{I}_V^k$. Nevertheless, solving this problem requires deep understanding of the behavior of these statistics as a function of Π, whichever model is assumed for the graph ensemble. For the remainder of our discussion, we shall assume it follows the $G_k(n, p, s)$ model.

6.3 Full mismatch

Our first attempt at determining feasibility results will consider the full edge mismatch statistic Δ_F. This statistic, as a function of Π, is intuitively connected to the "correctness" of Π, since the structural correlation induced by the generator graph over our graph ensemble implies that hyperpermutations that make few errors in matching vertices (and are, therefore, similar to $\mathbb{I}_V^k$) will observe a smaller amount of edge mismatches. A rather straightforward execution of this approach consists of analysing all hyperpermutations Π and bounding the probability that the estimator at hand is equal to Π, that is, the probability that Π minimizes Δ_F. The goal is to ensure that this probability, summed over all $\Pi \neq \mathbb{I}_V^k$, is vanishing.

Indeed, the general structure of the feasibility proofs in both [27] and [55], for the case of two graphs, follows this approach by taking the mismatch measure Δ and bounding the probability that each permutation has a value of Δ as great as the correct one. However, the combinatorial structure of the graph matching problem is relatively simple. Consider the following expression for $\Delta(\pi; G_1, G_2)$:

$$\Delta(\pi; G_1, G_2) = \sum_{e\in\binom{V}{2}} G_e^1 \otimes G_{\pi(e)}^2,$$

where $\otimes$ is the binary exclusive-or operator. Recall that G_e^i equals 1 if the edge set of G_i contains e, and equals 0 otherwise.

Note that, by the definition of the $G(n, p, s)$ model, each e-indexed term in this summation is independent of all but two other terms: the one indexed by $\pi(e)$ and

the one indexed by $\pi^{-1}(e)$. This cyclic dependence structure allows [27] to split this summation into a few sums of independent random variables, which allows for efficient applications of concentration results like the Chernoff bound.

This situation is much more complicated in the multiple graph matching problem. For instance, recall that:

$$\Delta_F = \sum_{i,j\in[k]} \sum_{e\in\binom{V}{2}} \mathbb{1}_{\{e\in E_i \otimes \Pi_{ij}(e)\in E_j\}}.$$

Each term in this expression is potentially correlated with 4 terms for each marginal permutation Π_{ij}, for a total of $4\binom{k}{2}$ terms.

Not only that, but the number of hyperpermutations that we need to analyze is equal to $(n!)^k$, which grows exponentially in k. This also makes the task of achieving meaningful bounds on the overall error probabilities much harder.

Our first result bypasses these difficulties by using the fact that $\mathbb{I}_V^k$ is the only hyperpermutation which has all marginals equal to $\mathbb{I}_V$. On the other hand, since it does not exploit the structural correlation of all graphs, it is unsurprisingly a result that gets weaker as k grows.

Lemma 6.6. *Let $G_1, \ldots, G_k$ be graphs distributed according to the $G_k(n,p,s)$ model, and let T be an arbitrary tree over $[k]$. If $p = o(1)$, s is constant and*

$$nps\frac{s^2}{32} = \log n + \frac{1}{2}\log k + \omega(1),$$

then $\mathbb{I}_V^k$ minimizes Δ_S and Δ_T a.a.s.

If, in addition,

$$nps\frac{s^2}{32} = \log n + \log k + \omega(1),$$

then it also minimizes Δ_F a.a.s.

Proof. For simplicity, let's consider only Δ_F on this proof. It is analogous to consider Δ_T, with minor adaptations on the argument, and Δ_S is a particular case of Δ_T, where T is a star.

Recall that:

$$\Delta_F = \sum_{i,j\in[k]} \Delta(\Pi_{ij}; G_i, G_j)$$

Fix any two graphs G_i and G_j in our input ensemble, and let A_{ij} be the number of non-identity permutations such that $\Delta(A_{ij}; G_i, G_j)$ is minimum. Under the

conditions on the hypothesis, it holds that[2]:

$$\mathbb{E}[A_{ij}] \leq 2\frac{\exp(2(\log n - nps\frac{s^2}{32}))}{1 - \exp(\log n - nps\frac{s^2}{32})}.$$

Since A_{ij} is an integer random variable, applying Markov's inequality yields:

$$\mathbb{P}[A_{ij} > 0] \leq 2\frac{\exp(2(\log n - nps\frac{s^2}{32}))}{1 - \exp(\log n - nps\frac{s^2}{32})}$$

Note that $\Delta(\Pi_{ii}; G_i, G_i)$ measures the mismatch between G_i and itself via $\Pi_{ii} = \mathbb{I}_V$, which is trivially equal to 0 for any Π. Therefore, we can ignore terms with $i = j$ on the expression for Δ_F. For the remaining terms, union bound implies:

$$\begin{aligned}
\mathbb{P}[\exists\ a \neq b\ :\ A_{ij} > 0] &\leq \sum_{a \neq b} \mathbb{P}[A_{ij} > 0] \\
&\leq 2\binom{k}{2}\frac{\exp(2(\log n - nps\frac{s^2}{32}))}{1 - \exp(\log n - nps\frac{s^2}{32})} \\
&= \frac{\exp(\log k + \log(k-1) + 2(\log n - nps\frac{s^2}{32}))}{1 - \exp(\log n - nps\frac{s^2}{32})} \\
&\leq \frac{\exp(2(\log k + \log n - nps\frac{s^2}{32}))}{1 - \exp(\log n - nps\frac{s^2}{32})}
\end{aligned} \tag{6.1}$$

Under the hypothesis of this lemma, the right hand side of this inequality vanishes, thus each term in the expression of Δ_F is minimized uniquely with $\Pi_{ij} = \mathbb{I}_V$, a.a.s. This already implies $\mathbb{I}_V^k$ is the unique minimum of Δ_F, a.a.s. □

As we have previously stated, while this result implies that having multiple graphs to match has at most a mild negative impact on our assumptions, it does not allow us to exploit the additional information available. One can argue that the proof of this result does not explore in full the technique used by Pedarsani and Grossglauser, instead reducing the multiple graph matching problem to a series of two-graph matching problems.

Let us explore their technique more carefully. It consists on a number of ingredients that build up to the following statements: if S is the random variable (function of the $G(n, p, s)$ random graphs) that counts how many permutations are as good as the identity $\mathbb{I}_V$ (that means, satisfy $\Delta(\pi) < \Delta(\mathbb{I}_V)$), then S converges to zero in expectation and, by the First Moment Method, must also converge to zero in

[2] A stronger version of this result has been shown in the course of the proof of Theorem 4.1 in [27]. However, the proof of this theorem as published is flawed. The statement we use is a corrected version extracted from an unpublished appendix of [57].

probability. This random variable S is a sum of indicators $\mathbb{1}_{\{\Delta(\pi)<\Delta(\mathbb{I}_V)\}}$, which runs through all non-identity permutations π. We'll call these the *goodness indicators*. To bound its expectation amounts, then, to bounding the corresponding *goodness probability* $\mathbb{P}[\Delta(\pi) < \Delta(\mathbb{I}_V)]$ of each permutation[3].

Some permutations are more similar to $\mathbb{I}_V$, and therefore are more likely to make a similar number of edge mismatches. Since the overall number of permutations is very large, uniformly upper-bounding the goodness of all permutations by these similar-to-$\mathbb{I}_V$ specimen might overestimate the goodness of too many permutations, and give us a too loose, not vanishing, upper bound.

Their solution to this problem is to split the permutations into classes, according to their similarity to $\mathbb{I}_V$ — more specifically, acoording to their *order*, the number of its non-fixed points. A permutation of order 2 is a simple transposition of two elements, while a permutation of order n "leaves no stone unturned". The goodness of the permutations can then be upper-bounded more tightly, and uniformly within each of these classes.

Now, as a function of the order, classes might get exponentially larger, while their goodness upper bound decrease at least exponentially. This allows for bounding $\mathbb{E}[S]$ by an infinite geometric series. The hypotheses of their result, ultimately, ensure that: (i) this geometric series is convergent, with ratio bounded away from 1 in absolute value and (ii) its first term, as a function of n, converge to 0. (i) implies the series is bounded by its first term times a constant, which coupled with (ii) gives the desired result.

The most complicated technical aspect is determining a good enough upper bound on the goodness probability of a permutation, $\mathbb{P}[\Delta(\pi) < \Delta(\mathbb{I}_V)]$. Several steps were taken to achieve this. First, it is noted that every π-invariant possible edge contributes nothing to $\Delta(\pi) - \Delta(\mathbb{I}_V)$, since either both permutations mismatch this possible edge or both match them correctly. This difference can then be rewritten to consider only possible edges that are not π-invariant (which are more numerous for permutations of higher order).

The rewritten expression can still be interpreted as some mismatch difference between π and $\mathbb{I}_V$, which is written in the proof as $X_\pi - Y_\pi$ (with X_π being mismatches by π and Y_π being mismatches by $\mathbb{I}_V$). However, we know every contributing possible edge is not π-invariant. This can be used to estimate the distribution of X_π and Y_π. More accurately, this allows to decompose the expression for X_π into sets of terms, such that terms in different sets are independent. This keeps the correlation under control and stochastically lower-bounds X_π by a binomial random variable, with some parameters. Y_π doesn't even need this aid: its definition depends only on

[3]Note that, in this expression, the permutations are not random; what is random is the value of Δ for both π and $\mathbb{I}_V$, since it depends on the $G(n, p, s)$ sample.

π and on the structure of $G(n, p, s)$, and it can be shown to be distributed precisely as a binomial random variable, with some other parameters.

Under mild conditions taken under hypothesis, the binomial for X_π has larger expected value than the one for Y_π. An additional lemma ensures that the first binomial will be larger than the second binomial, with probability that goes to 1 exponentially fast as a function of the "difference"[4] of their averages. All other parameters fixed, this difference is at least linear on the order of the permutation, which feeds into the framework we have just described. Most importantly, this lemma is powerful enough to not assume independence of the two binomials.

In extending this technique to the problem of multiple graph matching, we're required to deal simultaneously with a more complicated permutation set (which has grown from the symmetric group $\mathrm{Sym}(V)$ to the hypersymmetric group $\mathrm{Sym}_k(V)$) and a more complicated error function (such as Δ_S or Δ_T), with several correlated terms. Certainly, different ways of splitting the set of hyperpermutations will provide bounds of different kinds on the behavior of the error functions. Also, from the analysis we've just performed, any splitting of this set of hyperpermutations into classes should allow for the number of elements on each class to be counted (or at least upper-bounded) while ensuring a minimum number of matching "errors" to consider within these classes.

A feasible approach is to split our hyperpermutations by the order sequence of its marginals. This can be done as follows: let $\vec{m} = \{m_{ij}\}_{ij \in \binom{[k]}{2}}$ represent the desired order sequence ($\vec{m}$ being an element of $[0 : n]^{\binom{[k]}{2}}$), and denote by $\mathcal{C}_{\vec{m}}$ the class of hyperpermutations Π such that each of its marginals Π_{ij} has order m_{ij}. Note that every hyperpermutation belongs to a single class and, recalling that these marginals uniquely define a hyperpermutation, we can write

$$\mathrm{Sym}_k(V) = \biguplus_{\vec{m}} \mathcal{C}_{\vec{m}}$$

and obtain a full, disjoint partition scheme.

The cardinality of each class is bounded by $\prod_{ij \in \binom{[k]}{2}} R_{n,n-m_{ij}}$, where $R_{n,i}$ denotes the *rencontres numbers*[5]. It is known that $R_{n,i} = \binom{n}{i} R_{n-i,0}$ and that $R_{n-i,0} = [\frac{(n-i)!}{e}]$, where $[\cdot]$ is the ceiling function for even $n - i$ and the floor function for odd $n - i$. This implies $R_{n,i} \leq n^{n-i}$, since:

- if $n - i = 0$, $R_{n-i,0} = R_{0,0} = 1 \leq n^0 = n^{n-i}$;
- if $n - i = 1$, then $R_{n-i,1} = R_{1,0} = 0 \leq n = n^{n-i}$

[4] More precisely, on the ratio between their difference squared and their sum.

[5] The *rencontre number* $R_{n,i}$ counts all permutations over a set of size n with precisely i fixed points (that is, with order $n - i$).

- if $n - i > 1$, we have:

$$R_{n,i} \leq \binom{n}{i}\left(\frac{(n-i)!}{e} + 1\right) \leq n^{n-i}\left(\frac{1}{e} + \frac{1}{(n-i)!}\right) \leq n^{n-i},$$

since both e and $(n-i)!$ are greater or equal to 2.

Therefore, it holds that:

$$\prod_{ij \in \binom{[k]}{2}} R_{n,n-m_{ij}} \leq \prod_{ij \in \binom{[k]}{2}} n^{m_{ij}} = n^{\|\vec{m}\|_1}$$

where $\|\cdot\|_1$ denotes the vector 1-norm.

Now, it still remains to bound the "goodness probability" inside each of these classes. This requires us to understand the behavior of:

$$\Delta_S(\Pi) - \Delta_S(\mathbb{I}_V^k) = \sum_{e \in \binom{V}{2}} \sum_{ij \in \binom{[k]}{2}} \mathbb{1}_{\{e \in E_i \otimes \Pi_{ij}(e) \in E_j\}} - \mathbb{1}_{\{e \in E_i \otimes e \in E_j\}}$$

for Π in each class of hyperpermutations $\mathcal{C}_{\vec{m}}$.

We start by noting that, in the previous summation, several terms contribute equally to $\Delta_S(\Pi)$ and $\Delta_S(\mathbb{I}_V^k)$ regardless of the structure of our random graphs. This will happen whenever, for $ij \in \binom{[k]}{2}$ and $e \in \binom{V}{2}$, e is a *fixed pair*[6] of the marginal permutation Π_{ij}, since this implies the corresponding terms on our difference are the same random variable.

Denote by $R_\Pi \subseteq \binom{[k]}{2} \times \binom{V}{2}$ the indices of remaining terms, that is:

$$R_\Pi = \left\{(ij, e) \in \binom{[k]}{2} \times \binom{V}{2} \ : \ e \neq \Pi_{ij}(e)\right\}.$$

The size of this set will be denoted by $|R_\Pi| = r_\Pi$. Then, we can rewrite the summation as:

$$\Delta_S(\Pi) - \Delta_S(\mathbb{I}_V^k) = \sum_{(a,e) \in R_\Pi} \mathbb{1}_{\{e \in E_i \otimes \Pi_{i,j}(e) \in E_j\}} - \mathbb{1}_{\{e \in E_i \otimes e \in E_j\}}$$

We'll introduce the following notation to simplify our expressions. For $(ij, e) \in$

[6] We define a *fixed pair* of a function as an unordered pair in its domain which is the image of itself under such function. It must be either a pair of fixed points, or a pair of points such that one is the image of the other.

$\binom{[k]}{2} \times \binom{V}{2}$, let:

$$\begin{aligned} X_{(ij,e)} &= \mathbb{1}_{\{e \in E_i \otimes \Pi_{i,j}(e) \in E_j\}} \\ Y_{(ij,e)} &= \mathbb{1}_{\{e \in E_j \otimes e \in E_j\}} \\ X_\Pi &= \sum_{(ij,e) \in R_\Pi} X_{(ij,e)} \\ Y_\Pi &= \sum_{(ij,e) \in R_\Pi} Y_{(ij,e)} \end{aligned}$$

We'll also denote by $\mathcal{I} = \binom{[k]}{2} \times \binom{V}{2}$ the range of indices for $X_{(ij,e)}$ and $Y_{(ij,e)}$.

Each random variable in the family $\{X_{(ij,e)}\}$ is indexed by a pair of graphs G_i, G_j and an edge e, and indicates whether the edge e and its map $\Pi_{i,j}(e)$ appear consistently on graphs G_i and G_j. Note that this definition depends on the chosen hyperpermutation Π, but this relationship is omitted from our notation. The corresponding random variable $Y_{(ij,e)}$ checks the same consistency, but using the identity hyperpermutation to map an edge between G_i and G_j.

Our expression can then be shortened to:

$$\Delta_S(\Pi) - \Delta_S(\mathbb{I}_V^k) = \sum_{(ij,e) \in R_\Pi} X_{(ij,e)} - Y_{(ij,e)} = X_\Pi - Y_\Pi$$

Several of the remaining terms are still correlated, as the same edge e appears multiple times on this summation and the corresponding terms are correlated through the common generator graph. One way to deal with this is to split these terms into several groups, such that all terms that belong to the same group are mutually independent. This way, we can handle non-independent variables by an intergroup union bound, while applying tighter bounds on summations inside each group.

To do that, we'll need one more definition and three supporting lemmas.

Definition 6.7. *Two indices $(a, e), (b, e') \in \mathcal{I}$ satisfy the* disjoint edge set *criterion for Π if the sets*

$$\{e\ ,\ \Pi_a(e)\} \quad , \quad \{e'\ ,\ \Pi_b(e')\}$$

are disjoint.

Note that, in this definition, a and b denote pairs of graphs. The idea behind this definition is that the two sets just described can be used to identify more basic random variables used to construct our random variables $X_{(a,e)}$ and $X_{(b,e')}$. These sets being disjoint informally states that $X_{(a,e)}$ and $X_{(b,e')}$ are defined by different random variables, which leads to their independence (similarly for $Y_{(a,e)}$ and $Y_{(b,e')}$). This is precisely what our first supporting lemma shows.

Lemma 6.8. *Consider two indices $(a, e), (b, e') \in \mathcal{I}$. If they satisfy the disjoint edge set criterion for Π, then:*

- *$X_{a,e}$ and $X_{b,e'}$ are pairwise independent.*
- *$Y_{a,e}$ and $Y_{b,e'}$ are pairwise independent.*

Note that this criteria is not necessarily exhaustive and this family might have even more pairwise independent random variables. However, identifying these pairs will suffice for now.

We omit the proof of this lemma and, in fact, it will not be used directly, but it serves as an introduction for our second lemma, as this reasoning will be extended. Now, instead of picking just two indices, assume we have a set of indices, and they all pairwise satisfy the disjoint edge set criterion. The previous lemma ensures that the corresponding sets of random variables are pairwise independent. However, we cannot yet assert that they are (mutually) independent, as this condition is stronger. The next lemma provides this conclusion:

Lemma 6.9. *For any subset of indices $R \subseteq \mathcal{I}$, if all indices in R pairwise satisfy the disjoint edge set criterion for Π, then:*

- *$\{X_{(a,e)}\}_{(a,e)\in R}$ are independent random variables.*
- *$\{Y_{(a,e)}\}_{(a,e)\in R}$ are independent random variables.*

Proof. We begin by noting that, for any subset $E \subseteq \binom{V}{2}$, for any choice of graphs $i(e)$, $e \in E$, the family of σ-algebras $\{\sigma(G_e^{i(e)})\}$ is an independent family. Now, for $a = ij$, $X_{(a,e)}$ is a function of the random variables G_e^i and $G^j_{\Pi_a(e)}$, therefore it is measurable on the σ-algebra $\sigma(G_e^i, G^j_{\Pi_a(e)})$. Denote this σ-algebra by $\sigma_{(a,e)}$.

If R satisfies the disjoint edge set criterion for Π, there are no variable G_e^i which generates more than one σ-algebra in the family $\{\sigma_{(a,e)}\}_{(a,e)\in R}$. This implies this family is generated by independent families of random variables, and thereby form an independent family of σ-algebras. We conclude that the random variables $\{X_{(a,e)}\}_{(a,e)\in R}$ form an independent family, which is our first conclusion.

The second conclusion follows analogously. □

The third lemma ensures there is at least one way to group these random variables as we described before, such that we have a small number of similarly-sized groups.

Lemma 6.10. *For any hyperpermutation Π, there exist an integer $l \leq 4\binom{k}{2} - 1$ and a partition $R_\Pi = R_1 \uplus \cdots \uplus R_l$ such that:*

1. *$r_i := |R_i| \in \{\lfloor r_\Pi/l \rfloor, \lceil r_\Pi/l \rceil\}$ for all $i \leq l$;*

2. *for all* $i \leq l$*, the random variables* $\{X_{(a,e)}\}_{(a,e)\in R_i}$ *are independent;*

3. *for all* $i \leq l$*, the random variables* $\{Y_{(a,e)}\}_{(a,e)\in R_i}$ *are independent.*

It is worth commenting on the principle behind the proof beforehand, as it can be used in completely different contexts. The main scenario is that we have a family of random variables (in our case, indexed by a convenient set) with some convoluted independence pattern among them. As already mentioned, our goal is to partition this family into subsets of independent random variables. The approach is to arrange these random variables as vertices of a graph, such that the absence of an edge implies independence of the corresponding random variables. We'll call this graph the *dependence graph*[7]. Any independent set on this graph corresponds to a set of independent random variables, so the desired partition is equivalent to a vertex coloring of this graph, with each subset being extracted from a different color class.

If one is interested in minimizing the number of subsets in the partition, then the desired coloring is one that realizes the chromatic number of the dependence graph. On the other hand, if the sizes of each subset also matter (for example, if Chernoff-like bounds are to be applied on each subset), then one might be more interested in finding an *equitable coloring*, in which all color classes are as equal-sized as possible. The latter is certainly our case, and we'll use a graph-theorical result called the *Hajnal-Szemerédi* theorem to construct our equitable coloring of the dependence graph. This is not a novel approach [63–65], but it is nevertheless an interesting tool that we believe has not received its fair share of attention.

In applying this approach, it is certainly necessary to understand the graph-theoretical features of the dependence graph. In general, one would like this graph to be as sparse as possible, as this aids in minimizing its chromatic number and obtaining a smaller number of larger color classes. However, in most contexts, one has to ensure that the pairwise independence in each subset implies its mutual independence. To achieve this, it might be necessary to be conservative about which "kinds of independence" will be considered upon building the dependence graph, which might lead one to settle for a denser dependence graph and, thus, a worse partition of the family of random variables. This is not an issue in our case by virtue of Lemma 6.9.

We can now proceed to the proof of Lemma 6.10.

Proof. Denote by DG the dependence graph for the hyperpermutation Π, having as node set our set R_Π of remaining terms, and as edge set all pairs $(a,e),(b,e') \in R_\Pi$

[7]This arrangement of random variables in the form of a graph is similar to Markov random fields, even though the independence properties that hold in such a setting are distinct from the independence properties obtained for our dependence graph.

that don't satisfy the disjoint edge set criterion. Following standard graph theory notation, denote by Δ the maximum degree in DG. Then, by the Hajnal-Szemerédi theorem, DG has an equitable coloring that uses precisely $l = \Delta + 1$ colors.

Denote by $R_1, \ldots, R_l$ the corresponding disjoint edge classes. Since our coloring is equitable, each color class has either size $\lfloor r_\Pi / l \rfloor$ or $\lceil r_\Pi / l \rceil$. Also, each color class R_i is an independent set and, by construction, its nodes pairwise satisfy the disjoint edge set criterion. By Lemma 6.9, the corresponding random variables $X_{(\cdot,\cdot)}$ are independent, and so are the random variables $Y_{(\cdot,\cdot)}$.

Finally, to upper-bound l, note that any node (a, e) can be adjacent in DG to, at most, $4\binom{k}{2} - 2$ other nodes (as long as they are elements of R_Π):

- 2 nodes of the form (a, e'), satisfying either $e' = \Pi_a(e)$ or $e = \Pi_a(e')$;
- for each $b \neq a$ ($\binom{k}{2} - 1$ such values), 4 nodes (b, e') satisfying one of the following: $e' = e$; $e' = \Pi_a(e)$; $e = \Pi_b(e')$; or $\Pi_a(e) = \Pi_b(e')$, for a total of $4\binom{k}{2} - 4$ nodes.

Since all degrees are bounded by $4\binom{k}{2} - 2$, so is Δ, which implies $l \leq 4\binom{k}{2} - 1$. $\square$

Now we can proceed to apply our decomposition procedure, which culminates in the following result:

Lemma 6.11. *For any fixed hyperpermutation Π, let $R_\Pi = \left\{(ij, e) \in \binom{[k]}{2} \times \binom{V}{2} \ : \ e \neq \Pi_{ij}(e)\right\}$ and $r_\Pi = |R_\Pi|$. Then,*

$$\mathbb{P}[\Delta_F(\Pi) - \Delta_F(\mathbb{I}_V^k) \leq 0] \leq 2 \exp\left\{\log 4 \left(\binom{k}{2} - 1\right) - \left(\frac{r_\Pi}{4\binom{k}{2} - 1} - 1\right) \cdot \frac{ps((1-p)s)^2}{6(1-ps)}\right\}$$

Proof. Recall, from the notation we introduced earlier, that $\Delta_F(\Pi) - \Delta_F(\mathbb{I}_V^k) = X_\Pi - Y_\Pi$. Using l and $R_1, \ldots, R_l$ from Lemma 6.10, for $1 \leq l' \leq l$, define:

$$X_\Pi(R_{l'}) = \sum_{(a,e) \in R_i} X_{(a,e)},$$
$$Y_\Pi(R_{l'}) = \sum_{(a,e) \in R_i} Y_{(a,e)}.$$

By the definition of X_Π and Y_Π:

$$X_\Pi = \sum_{l' \leq l} X_\Pi(R_{l'}),$$
$$Y_\Pi = \sum_{l' \leq l} Y_\Pi(R_{l'}).$$

All random variables $X_{(a,e)}$ indexed by $R_{l'}$ are independent, by virtue of Lemma 6.10. Moreover, denoting $a = ij$, each $X_{(a,e)}$ is a Bernoulli random variable, with $\mathbb{E}[X_{(a,e)}] = \mathbb{P}[e \in E_i \otimes \Pi_{ij}(e) \in E_j]$. Since $e \neq \Pi_{ij}(e)$ by the definition of R_Π, the events $\{e \in E_i\}$ and $\{\Pi_{ij}(e) \in E_j\}$ are independent and happen with probability ps. This implies $\mathbb{E}[X_{(a,e)}] \overset{d}{\sim} \text{Be}(2ps(1-ps))$, and therefore:

$$X_\Pi(R_{l'}) = \text{Bin}(r_{l'}, 2ps(1-ps)).$$

For $Y_{(a,e)}$, we note that the events $\{e \in E_i\}$ and $\{e \in E_j\}$ are not independent, but are conditionally independent given $\{e \in E\}$, and have probability 0 given $\{e \notin E\}$. This means $\mathbb{E}[Y_{(a,e)}] = \mathbb{P}[e \in E_i \otimes e \in E_j] = 2ps(1-s)$ and:

$$Y_\Pi(R_{l'}) = \text{Bin}(r_{l'}, 2ps(1-s)).$$

Denote the averages of there random variables by:

$$\mu_{l'} = \mathbb{E}[X_\Pi(R_{l'})] = 2r_{l'}ps(1-ps),$$
$$\nu_{l'} = \mathbb{E}[Y_\Pi(R_{l'})] = 2r_{l'}ps(1-s).$$

Taking any $\delta, \varepsilon \in (0,1)$, the following Chernoff bounds apply:

$$\mathbb{P}[X_\Pi(R_{l'}) \leq (1-\delta)\mu_{l'}] \leq \left\{ -\frac{\delta^2 \mu_{l'}}{2} \right\}$$
$$\mathbb{P}[Y_\Pi(R_{l'}) \geq (1+\varepsilon)\nu_{l'}] \leq \exp\left\{ -\frac{\varepsilon^2 \nu_{l'}}{3} \right\}$$

Denoting $\mu = \mathbb{E}[X_\Pi]$ and $\nu = \mathbb{E}[Y_\Pi]$:

$$\mu = \sum_{i \leq l} \mu_i = \left(\sum_{i \leq l} r_i \right) 2ps(1-ps) = 2r_\Pi ps(1-ps)$$
$$\nu = \sum_{i \leq l} \nu_i = \left(\sum_{i \leq l} r_i \right) 2ps(1-s) = 2r_\Pi ps(1-s),$$

Now, we can obtain a concentration bound for X_Π via union bound:

$$\begin{aligned}
\mathbb{P}[X_\Pi \leq (1-\delta)\mu] &\leq \mathbb{P}[\exists\ l' \ :\ X_\Pi(R_{l'}) \leq (1-\delta)\mu_{l'}] \\
&\leq \sum_{l' \leq l} \mathbb{P}[X_\Pi(R_{l'}) \leq (1-\delta)\mu_{l'}] \\
&\leq \sum_{l' \leq l} \exp\left\{-\frac{\delta^2 \mu_{l'}}{2}\right\} \\
&\leq l \cdot \max_{l' \leq l} \exp\left\{-\frac{2\delta^2 r_{l'} ps(1-ps)}{2}\right\} \\
&= \exp\left\{\log l - (\min_{l' \leq l} r_{l'})\delta^2 ps(1-ps)\right\}.
\end{aligned}$$

Proceeding analogously for Y_Π, we obtain:

$$\mathbb{P}[Y_\Pi \geq (1+\varepsilon)\nu] \leq \exp\left\{\log l - \frac{2}{3}(\min_{l' \leq l} r_{l'})\varepsilon^2 ps(1-s)\right\}.$$

So far, all we have shown is that X_Π (resp. Y_Π) is concentrated to the right (resp. left) of its expected value. Note that $p < 1$ ensures that $\mathbb{E}[X_\Pi] > \mathbb{E}[Y_\Pi]$, so these inequalities should be enough to ensure that $X_\Pi - Y_\Pi > 0$. Let's make this more concrete. The next steps are nearly the same as the ones performed by Pedarsani and Grossglauser for matching two graphs.

By union bound, the following inequality holds for arbitrary $t \in \mathbb{R}$, and regardless of independence between X_Π and Y_Π:

$$\begin{aligned}
\mathbb{P}[X_\Pi - Y_\Pi \leq 0] &\leq \mathbb{P}[X_\Pi \leq t \cup Y_\Pi \geq t] \\
&\leq \mathbb{P}[X_\Pi \leq t] + \mathbb{P}[Y_\Pi \geq t].
\end{aligned}$$

Choosing $t = (\mu + \nu)/2$, and substituing our concentration inequalities (with $\delta = (\mu - \nu)/2\mu$ and $\varepsilon = (\mu - \nu)/2\nu$) yields:

$$\begin{aligned}
\mathbb{P}[X_\Pi - Y_\Pi \leq 0] &\leq \exp\left\{\log l - \frac{(\min_{l' \leq l} r_{l'})(\mu-\nu)^2 ps(1-ps)}{4\mu^2}\right\} \\
&\quad + \exp\left\{\log l - \frac{(\min_{l' \leq l} r_{l'})(\mu-\nu)^2 ps(1-s)}{6\nu^2}\right\} \\
&= \exp\left\{\log l - \frac{(\min_{l' \leq l} r_{l'})(\mu-\nu)^2}{4r_\Pi^2 ps(1-ps)}\right\} \\
&\quad + \exp\left\{\log l - \frac{(\min_{l' \leq l} r_{l'})(\mu-\nu)^2}{6r_\Pi^2 ps(1-s)}\right\} \\
&\leq 2\exp\left\{\log l - \frac{(\min_{l' \leq l} r_{l'})(\mu-\nu)^2}{6r_\Pi^2 ps(1-ps)}\right\}.
\end{aligned}$$

Since, by Lemma 6.10, all $r_{l'}$ satisfy $r_{l'} \geq r_\Pi/l - 1$, we have:

$$\begin{aligned}
\mathbb{P}[X_\Pi - Y_\Pi \leq 0] &\leq 2\exp\left\{\log l - \frac{(r_\Pi/l-1)(\mu-\nu)^2}{6r_\Pi^2 ps(1-ps)}\right\} \\
&= 2\exp\left\{\log l - \frac{(r_\Pi/l-1)(r_\Pi ps(1-p)s)^2}{6r_\Pi^2 ps(1-ps)}\right\} \\
&\leq 2\exp\left\{\log\left(4\binom{k}{2}-1\right) - \left(\frac{r_\Pi}{4\binom{k}{2}-1}-1\right)\cdot\frac{ps((1-p)s)^2}{6(1-ps)}\right\},
\end{aligned}$$

which is the desired result. □

This is as far as we can go for a specific hyperpermutation, as r_Π depends on our choice of hyperpermutation. However, if we know the order of all marginal permutations in Π, the following bounds apply:

Lemma 6.12. *If a hyperpermutation Π belongs to class $\mathcal{C}_{\vec{m}}$, then*

$$\|\vec{m}\|_1(n-1) - \|\vec{m}\|_2^2/2 \leq r_\Pi \leq \|\vec{m}\|_1(n-1/2) - \|\vec{m}\|_2^2/2,$$

where $\|\cdot\|_1$ and $\|\cdot\|_2$ are the traditional vector 1-norm and 2-norm, respectively.

Proof. Recall that $r_\Pi = |R_\Pi|$ counts the number of pairs (e,a) such that e is not a fixed pair of marginal permutation Π_a. Note that, for each (e,a), if both endpoints of e are fixed points of Π_a, then e is a fixed pair of Π_a. Each Π_a has $n-m_a$ fixed points, which leads to $\binom{n-m_a}{2}$ fixed pairs. On the other hand, any other fixed pair of Π_a must be a transposition, and Π_a can have, at most, $m_a/2$ transpositions. This means:

$$\begin{aligned}
\sum_{a\in[2:k]}\left(\binom{n}{2} - \binom{n-m_a}{2} - \frac{m_a}{2}\right) \leq r_\Pi \leq \sum_{a\in[2:k]}\left(\binom{n}{2} - \binom{n-m_a}{2}\right) \\
\sum_{a\in[2:k]}\left(\binom{m_a}{2} + m_a(n-m_a) - \frac{m_a}{2}\right) \leq r_\Pi \leq \sum_{a\in[2:k]}\left(\binom{m_a}{2} + m_a(n-m_a)\right) \\
\sum_{a\in[2:k]} m_a\left(\frac{m_a-1}{2} + (n-m_a) - \frac{1}{2}\right) \leq r_\Pi \leq \sum_{a\in[2:k]} m_a\left(\frac{m_a-1}{2} + (n-m_a)\right) \\
\sum_{a\in[2:k]}(m_a(n-1) - m_a^2/2) \leq r_\Pi \leq \sum_{a\in[2:k]}(m_a(n-1/2) - m_a^2/2) \\
\|\vec{m}\|_1(n-1) - \|\vec{m}\|_2^2/2 \leq r_\Pi \leq \|\vec{m}\|_1(n-1/2) - \|\vec{m}\|_2^2/2
\end{aligned}$$

□

The previous two lemmas directly imply the following statement:

Lemma 6.13. *Let Π be a hyperpermutation taken from class $\mathcal{C}_{\vec{m}}$. Then*

$$\mathbb{P}[\Delta_F(\Pi) - \Delta_F(\mathbb{I}_V^k) \leq 0] \leq 2\exp\left\{\log\left(4\binom{k}{2} - 1\right) - \left(\frac{\|\vec{m}\|_1(n-1) - \|\vec{m}\|_2^2/2}{4\binom{k}{2} - 1} - 1\right) \cdot \frac{ps((1-p)s)^2}{6(1-ps)}\right\}$$

At this point, we can state our main theorem of this section:

Theorem 6.14. *Let $G_1, \ldots, G_k$ be graphs distributed according to the $G_k(n,p,s)$ model. If $p = o(1)$ and*

$$\left(\frac{n}{2k} - 1\right)\frac{ps^3}{6} = (2k+1)\left(\log\binom{k}{2} + \log n\right) + \omega(1),$$

then $\mathbb{I}_V^k$ minimizes the error function Δ_F a.a.s.

Proof. Let $H = |\{\Pi \ : \ \Delta_F(\Pi) < \Delta_F(\mathbb{I}_V^k)\}|$. H is a non-negative integer random variable, so it suffices to show that $\mathbb{E}[H] \to 0$, since that implies $H = 0$ a.a.s. by the First Moment Method. Now, note that

$$\begin{aligned}\mathbb{E}[H] &= \sum_{\Pi \in \mathrm{Sym}_k(V)} \mathbb{P}[\Delta_S(\Pi) - \Delta_S(\mathbb{I}_V^k) \leq 0] \\ &= \sum_{\vec{m}} \sum_{\Pi \in \mathcal{C}_{\vec{m}}} \mathbb{P}[\Delta_S(\Pi) - \Delta_S(\mathbb{I}_V^k) \leq 0] \\ &\leq \sum_{\vec{m}} |\mathcal{C}_{\vec{m}}| \max_{\Pi \in \mathcal{C}_{\vec{m}}} \mathbb{P}[\Delta_S(\Pi) - \Delta_S(\mathbb{I}_V^k) \leq 0],\end{aligned}$$

where $\vec{m}$ runs through all order vectors for hyperpermutations $\Pi \neq \mathbb{I}_V^k$. This implies:

- $\|\vec{m}\|_1 \geq 2(k-1)$, since any hyperpermutation other than $\mathbb{I}_V^k$ has at least $k-1$ marginal permutations of order 2 or higher;
- $\|\vec{m}\|_2^2 \geq 2\|\vec{m}\|_1$, since every element of $\vec{m}$ is a non-negative integer different from 1, and every such integer x satisfies $x^2 \geq 2x$;
- $\|\vec{m}\|_2^2 \leq n\|\vec{m}\|_1$, since all elements of $\vec{m}$ are smaller or equal to n.

Using these facts and Lemma 6.13:

$$\begin{aligned}
\mathbb{E}[H] &\leq \sum_{\vec{m}} n^{\|\vec{m}\|_1} \max_{\Pi \in \mathcal{C}_{\vec{m}}} \left[2 \exp \left\{ \log \left(4 \binom{k}{2} - 1 \right) - \right. \right. \\
&\qquad \left. \left. \left(\frac{\|\vec{m}\|_1 (n-1) - \|\vec{m}\|_2^2/2}{4\binom{k}{2} - 1} - 1 \right) \cdot \frac{ps((1-p)s)^2}{6(1-ps)} \right\} \right] \\
&\leq 2 \sum_{\vec{m}} \exp \left\{ \|\vec{m}\|_1 \log n + \log \left(4 \binom{k}{2} - 1 \right) - \right. \\
&\qquad \left. \left(\frac{\|\vec{m}\|_1 (n-1) - n\|\vec{m}\|_1/2}{4\binom{k}{2} - 1} - 1 \right) \cdot \frac{ps((1-p)s)^2}{6(1-ps)} \right\} \\
&\leq 2 \sum_{m=2(k-1)}^{nk(k-1)/2} \exp \left\{ m \log n + \log \left(4 \binom{k}{2} - 1 \right) - \right. \\
&\qquad \left. \left(\frac{m(n-1) - nm/2}{4\binom{k}{2} - 1} - 1 \right) \cdot \frac{ps((1-p)s)^2}{6(1-ps)} \right\}
\end{aligned}$$

Now, there are at most $\binom{k}{2}^m$ order vectors $\vec{m}$ with 1-norm m, since all coordinates must be non-negative integers different from 1. Thus:

$$
\begin{aligned}
\mathbb{E}[H] &\leq 2 \sum_{m=2(k-1)}^{nk(k-1)/2} \binom{k}{2}^m \exp\left\{ m \log n + \log\left(4\binom{k}{2} - 1\right) - \right. \\
&\qquad \left. \left(\frac{m(n-1) - nm/2}{4\binom{k}{2} - 1} - 1 \right) \cdot \frac{ps((1-p)s)^2}{6(1-ps)} \right\} \\
&\leq 2 \sum_{m=2(k-1)}^{\infty} \exp\left\{ m \log \binom{k}{2} + m \log n + \log\left(4\binom{k}{2} - 1\right) - \right. \\
&\qquad \left. \left(\frac{m(n-1) - nm/2}{4\binom{k}{2} - 1} - 1 \right) \cdot \frac{ps((1-p)s)^2}{6(1-ps)} \right\} \\
&\leq 2 \sum_{m=2(k-1)}^{\infty} \exp\left\{ \log\left(4\binom{k}{2} - 1\right) + \frac{ps((1-p)s)^2}{6(1-ps)} + \right. \\
&\qquad \left. m \left[\log \binom{k}{2} + \log n - \frac{n/2-1}{4\binom{k}{2} - 1} \cdot \frac{ps((1-p)s)^2}{6(1-ps)} \right] \right\} \\
&= 2 \exp\left\{ \log\left(4\binom{k}{2} - 1\right) + \frac{ps[(1-p)s]^2}{6(1-ps)} + \right. \\
&\qquad \left. 2(k-1) \left[\log \binom{k}{2} + \log n - \frac{n/2-1}{4\binom{k}{2} - 1} \cdot \frac{ps[(1-p)s]^2}{6(1-ps)} \right] \right\} \\
&\qquad \cdot \frac{1}{1 - \exp\left\{ \log \binom{k}{2} + \log n - \frac{n/2-1}{4\binom{k}{2}-1} \cdot \frac{ps[(1-p)s]^2}{6(1-ps)} \right\}}
\end{aligned}
$$

where the last equality requires the geometric series to be convergent. This happens if and only if

$$
\frac{n/2-1}{4\binom{k}{2} - 1} \cdot \frac{ps[(1-p)s]^2}{6(1-ps)} > \log \binom{k}{2} + \log n,
$$

which holds for n large enough under our hypothesis.

Finally, note that the second factor in the resulting product is, under these conditions, upper-bounded by a constant, and the first factor vanishes under our hypothesis. This proves our result. □

It is worth noting that, in the case $k = 2$, this generalization is able to recover the original result by Pedarsani and Grossglauser, except for a multiplicative constant. However, for general $k = k(n)$, despite the complexity of its proof, this result is asymptotically weaker than Lemma 6.6. We strongly believe this to be due to the very loose bounds used in enumerating hyperpermutations by the order vector of their marginals, which should be a focus point of improvement.

Nevertheless, our intermediate results leading to the main theorem itself, up to

Lemma 6.13, have been mostly tailored to the specifics of each hyperpermutation and built on tight bounds. For instance, Lemma 6.10 allows us to split all terms in the expression for $\Delta_F(\Pi) - \Delta_F(\mathbb{I}_V^k)$ into up to $4\binom{k}{2}$ groups, such that there is full independence for terms in each group. This bound is tight up to a constant, since every hyperpermutation's expression cannot yield a splitting of this kind into less than $\binom{k}{2}$ groups, one for each pair of graphs, hopefully spanning all possible edges $e \in \binom{V}{2}$ — indeed, this is the case only for $\mathbb{I}_V^k$. This leads us to believe that these results should be very valuable in future analyses.

6.4 Maximum likelihood estimation

All results that have been presented so far considered the full edge mismatch measure Δ_F. As we have seen, this measure can be straightforwardly interpreted as the sum of the edge measure Δ — proposed for the two graph matching problem — over all pairs of graphs. As a consequence, Δ_F is oblivious, in a sense, to the specific patterns used to map each vertex across all graphs. In this, we will explore an alternative mismatch statistic which we hope can help leverage the joint structure of the observed graphs in order to derive better bounds for feasibility of matching.

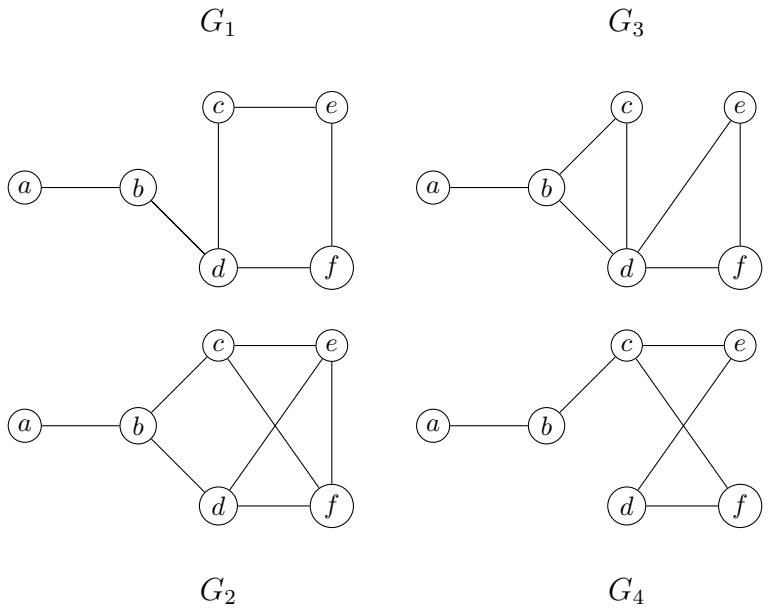
To begin, recall from section 6.1 that the *horizontal mapping* $\mathrm{Hor}_\Pi(v)$ of a vertex v by a hyperpermutation Π is the sequence of vertices given by $\{\Pi_{1j}(v)\}_{j\in[k]}$, that is, the sequence of all vertices which vertex v (in graph G_1) is mapped to in each graph G_j. The horizontal mapping of a vertex pair $e \in \binom{V}{2}$ is given similarly.

Definition 6.15. *Let $(G_1, \ldots, G_k)$ be k random graphs. For any $e \in \binom{V}{2}$, and any hyperpermutation Π of V, the* multiplicity *of e by Π, denoted $\mu_\Pi(e)$, is given by:*

$$\mu_\Pi(e) = \sum_{j\in[k]} \mathbb{1}_{\{[\mathrm{Hor}_\Pi(e)]_j \in E_j\}}.$$

To determine the multiplicity of a pair of vertices e by a hyperpermutation Π, we proceed as follows. Start by locating e in G_1. Then, for each graph $G_j, j \in [k]$, determine the pair of vertices to which e is mapped in G_j. This is given by the j-th position of the horizontal mapping of e. Call this edge e_j. Finally, for each e_j, count 1 if e_j is an edge of E_j, and 0 otherwise, and sum this quantity across all graphs. Figure 6.2 contains an illustration of this statistic.

Note that, while the horizontal mapping of e by Π is deterministic (it only depends on the chosen hyperpermutation Π), the multiplicity of e depends on the structures of all k graphs and is, therefore, a random variable. Furthermore, the sum of all multiplicities is independent of the chosen hyperpermutation Π and is given by $\sum_{e\in\binom{V}{2}} \mu_\Pi(e) = \sum_{j\in[k]} |E_j|$. This can be seen via a double counting argument: the



(a) 4 input graphs over a common vertex set $V = \{a, b, c, d, e, f\}$.

G_1	G_2	G_3	G_4
a	c	f	a
b	d	d	f
c	a	e	c
d	e	c	b
e	b	b	e
f	f	a	d

(b) A hyperpermutation Π given in matrix form.

	G_1	G_2	G_3	G_4	
$\mathrm{Hor}_\Pi(ad)$	ad	ce	cf	ab	
$\mathbb{1}_{[\mathrm{Hor}_\Pi(a)]_j \in G_j}$	0	1	0	1	$\mu_\Pi(a) = 2$
$\mathrm{Hor}_\Pi(cf)$	cf	af	ae	cd	
$\mathbb{1}_{[\mathrm{Hor}_\Pi(cf)]_j \in G_j}$	0	0	0	0	$\mu_\Pi(a) = 0$

(c) Calculating multiplicities of edges ad and cf by Π.

Figure 6.2: Illustration of the process of calculating a given edge's multiplicity by a hyperpermutation Π. Note that this calculation depends on Π and its result also depends on which graph ensemble has been observed. Edge cf has multiplicity 0 by Π: we'll say that Π *neutralizes* cf.

left-hand side expression counts all edges in all observed graphs by running through all vertex pairs and summing their multiplicities, and the right-hand side does the same by running through all graphs and summing the number of edges in each of them.

In the event that e has multiplicity 0 by Π, we will say that Π *neutralizes* e. The (random) set of potential edges which Π neutralizes will be called the *neutral set* of Π and denoted by $N(\Pi) = N(\Pi; G_1, \ldots, G_k)$.

While the connection between the neutral set and the structure of hyperpermutations may not be obvious in a first glance, it is an important one which can be easily noticed by considering the inference version of the $G_k(n, p, s)$ model. This version is formulated by incorporating the label shuffling into the construction pro-

cess: after the sampling process, we pick a fixed hyperpermutation $\Pi \in \mathrm{Sym}_k(V)$ and apply it to the vertex sets of our graphs, by taking the first graph as a baseline and shuffling the remaining labels using the marginal permutations Π_{1j}. The edge sets of all graphs will be given by:

$$E_j = \left\{ \Pi_{1j}(e) \ : \ e \in \binom{V}{2}, G_e^j = 1 \right\}.$$

The connection between this model and the neutral set of a hyperpermutation is given by the following theorem:

Theorem 6.16. *Consider the inference version of the $G_k(n, p, s)$, and denote by Π_{MLE} the maximum likelihood estimator for Π. Then,*

$$\Pi_{MLE} = \arg\max_{\Pi} |N(\Pi)|.$$

This result can be interpreted as stating that the "best guess" for mapping multiple networks, under no additional information, is achieved by perfectly grouping as many non-edges as possible. Note that, when taking $k = 2$, all potential edges have multiplicity 0, 1 or 2. Since the number of non-edges in all graphs is independent of how we match the two graphs, maximizing the size of the neutral set is equivalent to minimizing the number of potential edges with multiplicity 1, which is precisely the edge mismatch between the two graphs. This means the MLE for the case of two graphs is, indeed, a particular case of the MLE for the general case.

Intuitively, we can expect the size of the neutral set to be correlated with the mismatch measures we have previously presented. Consider, for instance, the full edge mismatch Δ_F, which counts all edge mismatches between all pairs of graphs. Δ_F can be rewritten as follows:

$$\Delta_F(\Pi) = \sum_{e \in \binom{V}{2}} \sum_{i,j \in [k]} \mathbb{1}_{[\{\mathrm{Hor}_\Pi(e)]_i \in E_i \otimes [\mathrm{Hor}_\Pi(e)]_j \in E_j\}}.$$

Each term of this expression can be seen as an individual contribution from $\mathrm{Hor}_\Pi(e)$, the horizontal mapping of the potential edge e. This contribution is greater the closer $\mu_\Pi(e)$ is to $k/2$, when the horizontal mapping of e has roughly an equal number of edges and non-edges. Now, the sum of all multiplicities is equal to the total number of edges in all graphs, which makes it constant with respect to Π. This means that choosing Π to neutralize many potential edges forces other multiplicities to be large. For both kinds of potential edges (the ones neutralized and the ones with large multiplicity), their contributions to Δ_F must be small. In other words, we can expect hyperpermutations with large neutral sets N_Π to also have small values of Δ_F.

However, it is important to point out that the MLE statistic is based on horizontal mappings, which take into account how the marginal permutations of a given hyperpermutation are combined. This behavior contrasts with the structure of statistics like Δ_F, in which each marginal permutation is handled separately and all information regarding how they are related is ignored.

We will now proceed to the proof of Theorem 6.16.

Proof. This proof generalizes an unpublished proof by Kazemi, Hassani and Grossglauser for the case of 2 graphs. By default, we use neutral capitals to denote random objects and calligraphic capitals to denote deterministic objects, with the exception of V.

Let $\mathcal{L}(\Pi; \mathcal{G}_1, \dots \mathcal{G}_k)$ denote the likelihood function of Π given input graphs $\mathcal{G}_1 = (V, \mathcal{E}_1), \dots, \mathcal{G}_k = (V, \mathcal{E}_k)$. The definition of MLE implies

$$\Pi_{ML} = \arg\max_{\Pi} \mathcal{L}(\Pi; G_1, \dots, G_k) = \arg\max_{\Pi} \log \mathcal{L}(\Pi; G_1, \dots, G_k),$$

where the second identity follows from the logarithm being strictly increasing.

Now, the definition of likelihood function, for this version of the $G_k(n, p, s)$ model, states that

$$\mathcal{L}(\Pi; G_k, \dots, G_k) = \mathbb{P}_\Pi[G_1 = \mathcal{G}_1, \dots, G_k = \mathcal{G}_k],$$

where $\mathbb{P}_\Pi$ is the probability measure obtained when applying hyperpermutation Π. Furthermore, by the law of total probability,

$$\mathcal{L}(\Pi; \mathcal{G}_1, \dots, \mathcal{G}_k) = \sum_{\mathcal{G}} \mathbb{P}_\Pi[G_1 = \mathcal{G}_1, \dots, G_k = \mathcal{G}_k, G = \mathcal{G}],$$

where $\mathcal{G} = (V, \mathcal{E})$ ranges over all graphs on node set V.

Now, the set $\mathcal{N}(\Pi) = N(\Pi; \mathcal{G}_1, \dots, \mathcal{G}_k)$ is the set of vertex pairs neutralized by Π, when we observe the graph ensemble $(\mathcal{G}_1, \dots, \mathcal{G}_k)$. Note that this set is not a function of our summation index $\mathcal{G}$. Denote by $\overline{\mathcal{N}(\Pi)} = \binom{V}{2} \setminus \mathcal{N}(\Pi)$ its complement.

If, for any vertex pair $e \in \overline{\mathcal{N}(\Pi)}$, it also holds that $e \notin \mathcal{E}$, then $\mathbb{P}_\Pi[G_1 = \mathcal{G}_1, \dots, G_k = \mathcal{G}_k, G = \mathcal{G}] = 0$, since it's impossible for an edge absent in G to be present in any of the observed graphs. Therefore, for any $\mathcal{E} \not\supseteq \overline{\mathcal{N}(\Pi)}$, it holds that $\mathbb{P}_\Pi[G_1 = \mathcal{G}_1, \dots, G_k = \mathcal{G}_k, G = \mathcal{G}] = 0$.

Consider now the case $\mathcal{E} \supseteq \overline{\mathcal{N}(\Pi)}$. The event $\{G_1 = \mathcal{G}_1, \dots, G_k = \mathcal{G}_k, G = \mathcal{G}\}$ is equivalent to the following conditions (which we will describe textually to avoid too cumbersome notation):

- the generator graph must be precisely $\mathcal{G}$ — this will happen with probability $p^{|\mathcal{E}|}(1-p)^{\binom{n}{2}-|\mathcal{E}|}$, since G is an Erdős-Rényi random graph;

- for each edge $e \in \mathcal{E}$, it must be further sampled in the observed graphs G_j such that $\Pi_{1j}(e) \in \mathcal{E}_j$, and it must not be sampled in the observed graphs G_j such that $\Pi_{1j}(e) \notin \mathcal{E}_j$ — which happens with further probability $s^{\mu_\Pi(e)}(1-s)^{k-\mu_\Pi(e)}$, independently for each edge, by construction;
- each edge $e' \notin \mathcal{E}$ must also meet this criterion; however, since (i) these edges can't be sampled in any observed graph, as they are not present in the generator, and (ii) we know that each such edge does not appear in any observed graph, due to being in $\mathcal{N}(\Pi)$, this criterion is always met.

Therefore, it holds that, if $\mathcal{E} \supseteq \overline{\mathcal{N}(\Pi)}$:

$$\begin{aligned}
&\mathbb{P}_\Pi[G_1 = \mathcal{G}_1, \ldots, G_k = \mathcal{G}_k, G = \mathcal{G}] \\
&= p^{|\mathcal{E}|}(1-p)^{\binom{n}{2}-|\mathcal{E}|} \prod_{e \in \mathcal{E}} s^{\mu_\Pi(e)}(1-s)^{k-\mu_\Pi(e)} \\
&= p^{|\mathcal{E}|}(1-p)^{\binom{n}{2}-|\mathcal{E}|} s^{\sum_{e\in\mathcal{E}} \mu_\Pi(e)}(1-s)^{k|\mathcal{E}|-\sum_{e\in\mathcal{E}} \mu_\Pi(e)}
\end{aligned}$$

Now, recall that $\sum_{e\in\binom{V}{2}} \mu_\Pi(e) = \sum_{j\in[k]} |E_j|$. On the event $\{G_1 = \mathcal{G}_1, \ldots, G_k = \mathcal{G}_k, G = \mathcal{G}\}$, this implies $\sum_{e\in\mathcal{E}} \mu_\Pi(e) = \sum_{j\in[k]} |\mathcal{E}_j|$ (terms for $e \notin \mathcal{E}$ have been dropped from the left-hand summation since $\mu_\Pi(e) = 0$ in this case). Denoting $\mathcal{A} = \mathcal{E} \setminus \overline{\mathcal{N}(\Pi)}$, it holds that:

$$\begin{aligned}
&\mathbb{P}_\Pi[G_1 = \mathcal{G}_1, \ldots, G_k = \mathcal{G}_k, G = \mathcal{G}] \\
&= \begin{cases} p^{|\mathcal{E}|}(1-p)^{\binom{n}{2}-|\mathcal{E}|} s^{\sum_{j\in[k]} |\mathcal{E}_j|}(1-s)^{k|\overline{\mathcal{N}(\Pi)}|+k|\mathcal{A}|-\sum_{j\in[k]} |\mathcal{E}_j|} & \text{, if } \mathcal{E} \supseteq \overline{\mathcal{N}(\Pi)} \\ 0 & \text{, otherwise} \end{cases}
\end{aligned}$$

Replacing this in the expression for $\mathcal{L}(\cdot)$ yields:

$$
\begin{aligned}
\mathcal{L}(\Pi; \mathcal{G}_1, \ldots, \mathcal{G}_k) &= \sum_{\mathcal{G}} \mathbb{P}_{\Pi}[G_1 = \mathcal{G}_1, \ldots, G_k = \mathcal{G}_k, G = \mathcal{G}] \\
&= \sum_{\mathcal{G} \,:\, \mathcal{E} \supseteq \overline{\mathcal{N}(\Pi)}} p^{|\mathcal{E}|}(1-p)^{\binom{n}{2} - |\mathcal{E}|} \cdot \\
&\qquad\qquad s^{\sum_{j \in [k]} |\mathcal{E}_j|}(1-s)^{k|\overline{\mathcal{N}(\Pi)}| + k|\mathcal{A}| - \sum_{j \in [k]} |\mathcal{E}_j|} \\
&= \sum_{\mathcal{A} \subseteq \mathcal{N}(\Pi)} p^{|\mathcal{A}| + |\overline{\mathcal{N}(\Pi)}|}(1-p)^{\binom{n}{2} - |\mathcal{A}| - |\overline{\mathcal{N}(\Pi)}|} \cdot \\
&\qquad\qquad s^{\sum_{j \in [k]} |\mathcal{E}_j|}(1-s)^{k|\overline{\mathcal{N}(\Pi)}| + k|\mathcal{A}| - \sum_{j \in [k]} |\mathcal{E}_j|} \\
&= \left(\frac{p(1-s)^k}{1-p}\right)^{|\overline{\mathcal{N}(\Pi)}|} \left(\frac{s}{1-s}\right)^{\sum_{j \in [k]} |\mathcal{E}_j|} (1-p)^{\binom{n}{2}} \\
&\qquad \cdot \sum_{\mathcal{A} \subseteq \mathcal{N}(\Pi)} \left(\frac{p(1-s)^k}{1-p}\right)^{|\mathcal{A}|}
\end{aligned}
$$

To solve the summation, note that there are precisely $\binom{|\mathcal{N}(\Pi)|}{a}$ subsets $\mathcal{A}$ of $\mathcal{N}(\Pi)$ such that $|\mathcal{A}| = a$. Using this remark and Newton's binomial theorem:

$$
\begin{aligned}
\mathcal{L}(\Pi; \mathcal{G}_1, \ldots, \mathcal{G}_k) &= \left(\frac{p(1-s)^k}{1-p}\right)^{|\overline{\mathcal{N}(\Pi)}|} \left(\frac{s}{1-s}\right)^{\sum_{j \in [k]} |\mathcal{E}_j|} (1-p)^{\binom{n}{2}} \\
&\qquad \cdot \sum_{a=0}^{|\mathcal{N}(\Pi)|} \binom{|\mathcal{N}(\Pi)|}{a} \left(\frac{p(1-s)^k}{1-p}\right)^{a} \\
&= \left(\frac{p(1-s)^k}{1-p}\right)^{\binom{n}{2} - |\mathcal{N}(\Pi)|} \left(\frac{s}{1-s}\right)^{\sum_{j \in [k]} |\mathcal{E}_j|} (1-p)^{\binom{n}{2}} \\
&\qquad \cdot \left(1 + \frac{p(1-s)^k}{1-p}\right)^{|\mathcal{N}(\Pi)|} \\
&= [p(1-s)^k]^{\binom{n}{2}} \left(\frac{s}{1-s}\right)^{\sum_{j \in [k]} |\mathcal{E}_j|} \left(1 + \frac{1-p}{p(1-s)^k}\right)^{|\mathcal{N}(\Pi)|}
\end{aligned}
$$

Substituting this into the expression for the MLE:

$$
\begin{aligned}
\Pi_{ML} &= \arg\max_{\Pi} \log \mathcal{L}(\Pi; G_1, \ldots, G_k) \\
&= \arg\max_{\Pi} \log \left\{ [p(1-s)^k]^{\binom{n}{2}} \left(\frac{s}{1-s}\right)^{\sum_{j\in[k]} |E_j|} \left(1 + \frac{1-p}{p(1-s)^k}\right)^{|N(\Pi)|} \right\} \\
&= \arg\max_{\Pi} \left\{ \binom{n}{2} (\log p + k \log(1-s)) + \sum_{j\in[k]} |E_j| \log\left(\frac{s}{1-s}\right) + \right. \\
&\qquad \left. |N(\Pi)| \log\left(1 + \frac{1-p}{p(1-s)^k}\right) \right\}
\end{aligned}
$$

The first term inside the arg max is constant and can be discarded. The second term, despite being a random variable, is a constant function of Π, since the total number of edges in the observed graphs does not depend on Π; it can be discarded as well. This yields:

$$
\Pi_{ML} = \arg\max_{\Pi} \left\{ |N(\Pi)| \log\left(1 + \frac{1-p}{p(1-s)^k}\right) \right\}
$$

Since $(1-p)/(p(1-s)^k)$ is positive, the logarithmic term is a positive multiplicative constant and can also be discarded. This leaves the desired result:

$$
\Pi_{ML} = \arg\max_{\Pi} |N(\Pi)|.
$$

□

While the role of the neutral set as a mismatch statistic is a very important one from an inferential point of view, this result still leaves us far from closure. It is still unknown whether the MLE is a consistent estimator, and answering this question requires us to understand the behavior of N_Π as a function of Π. Ideally, we would like to achieve a result similar to the ones regarding Δ_F — that is, to determine conditions under which $\mathbb{I}_V^k$ maximizes $|N_\Pi|$ a.a.s.

However, the relationship between the structure of a hyperpermutation and the contribution of each potential edge to N_Π is a complex combinatorial problem. To unravel a bit of this complexity, we will need an additional definition. A *partition* of an integer k is a multiset $\{k_1, \ldots, k_i\}$ of integers such that $k_1 + \cdots + k_i = k$.

Definition 6.17. *Let Π be a k-hyperpermutation over a set X and $x \in X$. The* horizontal profile *(or profile) of x by Π is the unique partition $\{k_1, \ldots, k_i\}$ of k such that there are distinct elements $x_1, \ldots, x_i \in X$ with x_j appearing precisely k_j times in $\mathrm{Hor}_\Pi(x)$.*

Recall that, in our original interpretation of a k-hyperpermutation, the horizontal

mapping of an element x enumerates all elements to which x is mapped in each copy of X. We can think of this horizontal mapping as splitting the k copies of X in equivalence classes, such that x is always mapped to the same element of X within each class. The horizontal profile is, then, a direct representation of the sizes of these classes, which is unique up to the enumeration order of classes and thus unique as a multiset. Refer to Figure 6.3 for an example.

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	horizontal profile
a	b	a	c	c	d	a	b	a	$\{4, 2, 2, 1\}$
b	c	b	b	b	b	b	c	b	$\{7, 2\}$
c	a	c	a	d	c	d	a	c	$\{4, 3, 2\}$
d	d	d	d	a	a	c	d	d	$\{6, 2, 1\}$

Figure 6.3: A 9-hyperpermutation over $X = \{a, b, c, d\}$ given in matrix form, where the last column shows the horizontal profile of each element. The horizontal mapping of a contains 4 copies of a, 2 of b, 2 of c and 1 of d, thus its horizontal profile is $\{4, 2, 2, 1\}$. The horizontal profiles of b, c and d are determined similarly.

In the multiple graph matching problem, horizontal profiles allow us to summarize the information contained in the horizontal mapping of vertices and edges, in a form that is still quite meaningful in terms of mismatch statistics. The following result relates the horizontal profile of an edge and its contribution to $|N_\Pi|$.

Lemma 6.18. *Let $G_1, \ldots, G_k$ be graphs distributed according to the $G_k(n, p, s)$ model. If $e \in \binom{V}{2}$ has horizontal profile $\mathcal{K} = \{k_1, \ldots, k_{|\mathcal{K}|}\}$ by a hyperpermutation Π, then*

$$\mathbb{P}[e \in N(\Pi)] = \prod_{j=1}^{|\mathcal{K}|} [1 - p + p(1 - s)^{k_j}].$$

Proof. Recall that $e \in N(\Pi)$ if and only if $\mu_\Pi(e) = 0$. By the definition of horizontal profile, there are vertex pairs $e_1, \ldots, e_{|\mathcal{K}|}$ such that e_i appears k_i times in $\mathrm{Hor}_\Pi(e)$. Then, by the definition of multiplicity:

$$\begin{aligned}
\mu_\Pi(e) &= \sum_{j \in [k]} \mathbb{1}_{\{[\mathrm{Hor}_\Pi(e)]_j \in E_j\}} \\
&= \sum_{i=1}^{|\mathcal{K}|} \left[\sum_{j \,:\, [\mathrm{Hor}_\Pi(e)]_j = e_i} \mathbb{1}_{\{e_i \in E_j\}} \right]
\end{aligned}$$

All terms of the outer summation are independent, as each term is function exclusively of random variables G_{e_i} and $S^j_{e_i}$ ($j \in [k]$). Therefore:

$$
\begin{aligned}
\mathbb{P}[\mu_\Pi(e) = 0] &= \prod_{i=1}^{|\mathcal{K}|} \mathbb{P}\left[\sum_{j \,:\, [\mathrm{Hor}_\Pi(e)]_j = e_i} \mathbb{1}_{\{e_i \in E_j\}} = 0\right] \\
&= \prod_{i=1}^{|\mathcal{K}|} \mathbb{P}\left[\sum_{j \,:\, [\mathrm{Hor}_\Pi(e)]_j = e_i} G_{e_i} S_{e_i}^j = 0\right] \\
&= \prod_{i=1}^{|\mathcal{K}|} \mathbb{P}\left[G_{e_i} \cdot \sum_{j \,:\, [\mathrm{Hor}_\Pi(e)]_j = e_i} S_{e_i}^j = 0\right] \\
&= \prod_{i=1}^{|\mathcal{K}|} \mathbb{P}\left[G_{e_i} = 0 \vee \sum_{j \,:\, [\mathrm{Hor}_\Pi(e)]_j = e_i} S_{e_i}^j = 0\right] \\
&= \prod_{i=1}^{|\mathcal{K}|} \left(\mathbb{P}[G_{e_i} = 0] + \mathbb{P}\left[G_{e_i} = 1, \sum_{j \,:\, [\mathrm{Hor}_\Pi(e)]_j = e_i} S_{e_i}^j = 0\right]\right) \\
&= \prod_{i=1}^{|\mathcal{K}|} (1 - p + p(1-s)^{k_j}),
\end{aligned}
$$

where the last equality follows from independence of G_e and S_e^j. □

This expression indicates that the probability that a potential edge is neutralized by Π decays exponentially with the size (number of classes) of its horizontal profile, since $p = o(1)$ implies the product is composed of $|\mathcal{K}|$ terms close to $(1-p)$. The following result states this more precisely.

Lemma 6.19. *Let $G_1, \ldots, G_k$ be graphs distributed according to the $G_k(n,p,s)$ model. If $e \in \binom{V}{2}$ has horizontal profile $\mathcal{K} = \{k_1, \ldots, k_{|\mathcal{K}|}\}$ by a hyperpermutation Π, then*

$$(1-p)^{|\mathcal{K}|} \leq \mathbb{P}[e \in N(\Pi)] \leq [c(1-p)]^{|\mathcal{K}|},$$

where $c = c(p,s) = \exp\{p(1-s)/(1-p)\}$.

Note that c does not depend on k or Π, and tends to 1 as a function of n under the usual assumption that $p \to 0$.

Proof. The lower bound follows trivially from dropping the terms in $p(1-s)^{k_j}$ from the final expression in Lemma 6.18. For the upper bound, working this expression

a bit further yields:

$$\begin{aligned}
\mathbb{P}[e \in N(\Pi)] &= \prod_{j=1}^{|\mathcal{K}|}(1-p+p(1-s)^{k_j}) \\
&= (1-p)^{|\mathcal{K}|}\prod_{j=1}^{|\mathcal{K}|}\left(1+\frac{p}{1-p}(1-s)^{k_j}\right) \\
&\leq (1-p)^{|\mathcal{K}|}\prod_{j=1}^{|\mathcal{K}|}\left(1+\frac{p}{1-p}(1-s)\right) \\
&= (1-p)^{|\mathcal{K}|}\left(1+\frac{p}{1-p}(1-s)\right)^{|\mathcal{K}|} \\
&\leq (1-p)^{|\mathcal{K}|}\exp\left\{\frac{p}{1-p}(1-s)\right\}^{|\mathcal{K}|} \\
&= (1-p)^{|\mathcal{K}|}\cdot c^{|\mathcal{K}|}.
\end{aligned}$$

□

One difficulty yet to be handled in accurately estimating $|N(\Pi)|$ is the dependence between contributions of different potential edges. It is easy to see that, whenever the horizontal mappings of e and e', elements of $\binom{V}{2}$, by Π have no common elements, then their contributions to $|N(\Pi)|$ are independent. For general hyperpermutations, however, it is not clear how to untangle the pattern of dependencies between all contributions.

Another serious difficulty on the combinatorial side of this problem is to relate the horizontal profiles of vertices by a hyperpermutation and the horizontal profiles of potential edges induced by them. While the horizontal mapping of a potential edge is uniquely determined by the horizontal mappings of its endpoints, the same cannot be said about the horizontal profiles. Consider, for instance, the example illustrated by the hyperpermutation Π in Figure 6.4. We can see that, even though vertices v_1, v_2 and v_3 have the same horizontal profile (namely, $\{3,3,3\}$), the horizontal profiles of edges v_1v_2 and v_1v_3 are fundamentally different, with v_1v_2 having the same profile as their endpoints and the profile of v_1v_3 equal to $\{1,1,\ldots,1\}$. Similar examples can be constructed for perfect square, arbitrary large values of k: taking two vertices whose horizontal profiles contain $\sqrt{k}$ copies of the element $\sqrt{k}$, the profile potential edge with endpoints on these two vertices could comprise $\sqrt{k}$ copies of $\sqrt{k}$, or k copies of 1, or several other possibilities.

At first glance, it seems like one could at least conclude that the profile of a potential edge must be a refinement[8] of the profiles of its endpoints. However, our

[8]In this context, we consider a partition $\mathcal{K}'$ of k to be a *refinement* of a partition $\mathcal{K}$ if the elements of $\mathcal{K}'$ can be grouped and summed to obtain $\mathcal{K}$. More precisely, $\mathcal{K}'$ is a refinement of $\mathcal{K}$

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
a	a	a	b	b	b	c	c	c
b	b	b	c	c	c	a	a	a
c	c	c	a	a	a	b	b	b
d	e	f	d	e	f	d	e	f
e	f	d	e	f	d	e	f	d
f	d	e	f	d	e	f	d	e
g	g	g	g	g	h	h	h	h
h	h	h	h	h	g	g	g	g
ab	ab	ab	bc	bc	bc	ca	ca	ca
ad	ae	af	bd	be	bf	cd	ce	cf
gh	gh	gh	gh	gh	gh	gh	gh	gh

Figure 6.4: A 9-hyperpermutation Π over vertex set $V = \{a, b, c, d, e, f, g, h\}$, and an excerpt of its induced hyperpermutation over $\binom{V}{2}$.

example also presents a counterexample to this conclusion. Both vertices v_4 and v_5 have profiles equal to $\{5, 4\}$, while the potential edge v_4v_5 has profile equal to $\{9\}$. This is due to v_4 and v_5 constituting a fixed pair according to several marginals of Π. Generalizing this reasoning, the horizontal profile of an edge could have cardinality as low as half the cardinality of the profiles of this endpoints.

6.5 Statistical approach

Our final contribution of this chapter is a statistical test to detect the identity hyperpermutation. Designed for the $G_k(n, p, s)$ model in mind, it is based on the following observation regarding multiplicities. Consider a potential edge $e \in \binom{V}{2}$ with horizontal profile equal to $\{k\}$ by some hyperpermutation (which we will omit from our notation for now). Such horizontal profile implies that e is correctly mapped across all graphs. The multiplicity of e can have two rather distinct conditional behaviors: if e is an edge of the generator graph G, $\mu(e)$ will have distribution $\text{Bin}(k, s)$, which heavily concentrates around ks, and if e is not an edge of G, $\mu(e)$ will be equal to 0.

This means that e is extremely unlikely to have small (i.e., not too close to ks) but positive multiplicity. More formally, if, for $0 < \alpha < 1$, e has multiplicity (by a hyperpermutation Π) either equal to 0 or greater than $ks(1 - \alpha)$, we will say e is *α-consistent* by Π. This is an event that depends not only on the chosen potential edge e, but also on the random graphs $G_1, \ldots, G_k$. Then, the following lemma holds:

Lemma 6.20. *Let $G_1, \ldots, G_k$ be graphs distributed according to the $G_k(n, p, s)$*

if $\mathcal{K}'$ can be constructed by taking a multiset union of partitions of elements of $\mathcal{K}$, one partition for each element, accounting for multiplicity.

model. If $e \in \binom{V}{2}$ *has horizontal profile* $\{k\}$ *by a hyperpermutation* Π*, then for any* $0 < \alpha < 1$*,*

$$\mathbb{P}[e \text{ is not } \alpha\text{-consistent by } \Pi] \leq p \exp\{-ks\alpha^2/2\}.$$

Proof. We begin by noting that, if e has horizontal profile $\{k\}$ by Π, then e must have horizontal profile $(e, e, \ldots, e)$ by Π, and its multiplicity by Π is given by:

$$\mu_\Pi(e) = G_e \sum_{j \in [k]} S_e^j.$$

If e is not an edge of G, then $\mu_\Pi(e) = 0$ and e is $\gamma/\sqrt{ks}$-consistent by definition. Furthermore, conditionally on e being an edge of G (which happens with probability p), $\mu_\Pi(e)$ has distribution $\text{Bin}(k, s)$. Applying the Chernoff bound yields:

$$\begin{aligned}
\mathbb{P}[e \text{ is not } \alpha\text{-consistent by } \Pi] &= \mathbb{P}[e \text{ is not } \alpha\text{-consistent by } \Pi, G_e = 0] \\
&\leq \mathbb{P}[\mu_\Pi(e) \leq ks(1 - \gamma/\sqrt{ks}), G_e = 0] \\
&\leq p \exp\{-ks\alpha^2/2\}
\end{aligned}$$

□

This conclusion strongly depends on its profile having elements of large size and small cardinality. As a contrast, we present the following extremal examples:

- If a potential edge e' has profile $\{k/2, k/2\}$, then with probability $2p(1-p)$ exactly one of the potential edges e' is mapped to is an edge of G. Conditionally on this event, $\mu(e')$ has distribution $\text{Bin}(k/2, s)$ and expected value $ks/2$;
- If e' has profile $\{1, \ldots, 1\}$, then $\mu(e')$ has distribution $\text{Bin}(k, ps)$ and expected value kps;
- If e' has profile $\{1, k-1\}$, then with probability $p(1-p)s$, $\mu(e')$ is precisely 1;

In all of these cases, there is an event which: (i) holds with probability at least $p(1-p)s$, and (ii) implies the multiplicity of e' doesn't follow the prescribed behavior of "0 or close to ks".

Now, recall that the identity hyperpermutation $\mathbb{I}_V^k$ is the only one such that all potential edges e have profile equal to $\{k\}$. Moreover, any other hyperpermutation must have at least two vertices with profile different from $\{k\}$, which imply at least $2(n-2)$ potential edges will also have profiles different from $\{k\}$. This means all non-identity hyperpermutations have a large number of edges that can

serve as evidence that they, indeed, are not $\mathbb{I}_V^k$. If we look at a hyperpermutation as a "population" of potential edges, each with its own horizontal profile, we can statistically test the properties of these potential edges looking for evidence that a given hyperpermutation is not $\mathbb{I}_V^k$.

Our statistical test is posed as follows. We are given random graphs $G_1, \ldots, G_k$ random graphs, sampled from the $G_k(n, p, s)$ model and an unknown hyperpermutation Π mapping the vertex sets of all graphs. The null hypothesis H_0 is that Π is equal to $\mathbb{I}_V^k$, and we attempt to reject this hypothesis in favor of an alternative hypothesis H_1 that Π is not $\mathbb{I}_V^k$. As a test statistic, let $C_\alpha = C_\alpha(\Pi)$ be the number of potential edges that are not α-consistent, where α is a tunable parameter of the test. We reject the null hypothesis whenever $C_\alpha > 0$, and reject it otherwise. We will call this the *consistency test for hyperpermutations* with parameter α.

As with every statistical hypothesis test, we must account for both Type I and Type II errors. Type I errors, or false positive errors, occurs when H_0 is true but is nonetheless rejected, that is, when $\mathbb{I}_V^k$ is misjudged to be some other hyperpermutation. Note that, intuitively, the probability of a Type I error decreases with α, as higher values of α impose looser restrictions on the observed multiplicities of potential edges. This is reflected on the following result, which uses Lemma 6.20 to characterize the probability of this type of error:

Lemma 6.21. *For the consistency test for hyperpermutations with parameter α, the probability of a Type I error is, at most, $\binom{n}{2} \exp\{-ks\alpha^2/2\}$.*

Proof. A Type I error occurs whenever $C_\alpha = C_\alpha(\mathbb{I}_V^k) > 0$, that is, at least one potential edge e is not α-consistent for $\mathbb{I}_V^k$. Recall that all e have horizontal profile $\{k\}$ by $\mathbb{I}_V^k$. Denoting by p_α the probability that one of these potential edges is not α-consistent, Lemma 6.20 states that $p_\alpha \leq p \exp\{-ks\alpha^2/2\}$.

Furthermore, every potential edge e has horizontal mapping by $\mathbb{I}_V^k$ equal to $(e, e, \ldots, e)$. This means the multiplicities for all e are independent, as each of them depend only on random variables G_e and S_e^j, for all $j \in [k]$. The contributions to all e to α are also independent, so C_α has distribution $\text{Bin}\left(\binom{n}{2}, p_\alpha\right)$, and:

$$\begin{aligned}
\mathbb{P}[C_\alpha = 0] &= (1 - p_\alpha)^{\binom{n}{2}} \\
&\geq (1 - p \exp\{-ks\alpha^2/2\})^{\binom{n}{2}} \\
&\geq 1 - \binom{n}{2} p \exp\{-ks\alpha^2/2\}
\end{aligned}$$

Since the probability of a Type I error is given by $\mathbb{P}[C_\alpha > 0] = 1 - \mathbb{P}[C_\alpha = 0]$,

this proves the result. □

A much harder task is determining the probability of Type II errors, or false negative errors. An error of this kind occurs when H_0 is false but we failed to reject it, which means the input hyperpermutation $\Pi \neq \mathbb{I}_V^k$ is misjudged to be $\mathbb{I}_V^k$. Estimates of the probability of this kind of error will strongly depend on the structure of the input hyperpermutation. Intuitively, the role of α on this probabilities is reversed: increasing α loosens the restriction of α-consistency and make it harder to detect inconsistency of potential edges and reject H_0, thus increasing the probability of a Type II error. An estimate of this probability can be given, under a few conditions, for a relatively simple class of hyperpermutations, the hypertranspositions:

Lemma 6.22. *Let $v_1, v_2 \in V$ be distinct vertices and $A \subset [k]$ be non-empty. Denote by $T = T(v_1, v_2, A)$ the hypertransposition that transposes vertices v_1 and v_2 between graph sets $\{G_j \ : \ j \in A\}$ and $\{G_j \ : \ j \in [k] \setminus A\}$. Assume that $|A| \leq k/2$, without loss of generality. If $\alpha \leq 1 - \frac{|A|}{ks}$, then for the consistency test for hyperpermutations with parameter α, the probability of a Type II error on input T is, at most,* $\exp\{-2(n-2)p(1-p)(1-(1-s)^{|A|})\}$.

Proof. In order to avoid a Type II error, we want to ensure that at least one potential edge will not be α-consistent. First, consider some $v \neq v_1, v_2$, and assume that $1 \in A$ — we will lift this assumption soon enough. The horizontal mapping of $v_1 v$ by T has $|A|$ copies of $v_1 v$ (for graphs $G_j, j \in A$) and $k - |A|$ copies to $v_2 v$ (for the remaining graphs). Let E_1 be the event where $v_1 v$ is an edge of the generator graph G but $v_2 v$ is not. Then, $\mathbb{P}[E_1] = p(1-p)$ and, conditionally on E_1, $\mu_T(v_1 v)$ has distribution $\text{Bin}(|A|, s)$. This allows us to write:

$$\mathbb{P}[v_1 v \text{ is } \alpha\text{-consistent}|E_1] = \mathbb{P}[\mu_T(v_1 v) = 0|E_1] + \mathbb{P}[\mu_T(v_1 v) > ks(1-\alpha)|E_1]$$

Note that, in E_1, $\mu_T(v_1 v)$ cannot be larger than $ks(1-\alpha)$, since it is at most $|A|$ and $ks(1-\alpha) \geq |A|$ by hypothesis. Thus, $\mathbb{P}[\mu_T(v_1 v) > ks(1-\alpha)|E_1] = 0$ and:

$$\mathbb{P}[v_1 v \text{ is } \alpha\text{-consistent}|E_1] = \mathbb{P}[\mu_T(v_1 v) = 0|E_1] = (1-s)^{|A|}$$

Now, let E_2 be the event where $v_2 v$ is an edge of G but $v_1 v$ is not. A similar calculation shows that $\mathbb{P}[E_2] = p(1-p)$ and $\mathbb{P}[v_2 v \text{ is } \alpha\text{-consistent}|E_2]$ is also upper-bounded by $(1-s)^{|A|}$. This implies that at least one of $v_1 v$ and $v_2 v$ is not α-consistent with probability at least $2p(1-p)(1-(1-s)^{|A|})$.

Note that, if $1 \notin A$, then $1 \in [k] \setminus A$ and we can reach the same conclusion by an analogous argument using $\mathbb{P}[v_1 v \text{ is } \alpha\text{-consistent}|E_2]$ and $\mathbb{P}[v_2 v \text{ is } \alpha\text{-consistent}|E_1]$. This allows us to lift the assumption that $1 \in A$.

Moving on, we note that the pair of events $\{v_1 v \text{ is } \alpha\text{-consistent}\}$ and $\{v_2 v \text{ is } \alpha\text{-consistent}\}$ depends only on random variables G_e and S_e^j, $j \in k$ for $e = v_1 v$ or $e = v_2 v$. As a consequence, these pairs of events, indexed by $v \in V \setminus \{v_1, v_2\}$, are mutually independent, since they are measurable in independent σ-algebras. Using this fact:

$$\begin{aligned}
\mathbb{P}[\text{Type II error}] &\leq \mathbb{P}[v_1 v', v_2 v' \text{ are } \alpha\text{-consistent } \forall\ v' \neq v_1, v_2] \\
&= (\mathbb{P}[v_1 v, v_2 v \text{ are } \alpha\text{-consistent}])^{n-2} \\
&= [1 - 2p(1-p)(1-(1-s)^{|A|})]^{n-2} \\
&\leq \exp\{-2(n-2)p(1-p)(1-(1-s)^{|A|})\}
\end{aligned}$$

□

Lemma 6.22 states that, for a hypertransposition, the probability of a Type II error decays exponentially fast as with the average degree of the input graphs, with the actual exponential bound being tighter when $|A|$ is larger. Note that, for this simple class of hyperpermutations, we were able to obtain interesting bounds that only slightly depend on α. While this is not expected to hold more generally, we believe this proof is a stepping stone towards generalizing this analysis to any hyperpermutation, thus precisely characterizing the probability of a Type II error and allowing for an appropriate choice for α.

Chapter 7

Conclusion and future work

"Like your father once said
Life is not what you're given
It is how you decide to live
On the path you have chosen"
(John Petrucci)

In this thesis, we addressed two topics on identity and structure in random networks. The first part was dedicated to the study of local symmetry, an analytical technique for studying structural equivalence of localities in networks. We have introduced the concept of *local symmetry*, the existence of isomorphism between neighborhoods of a given radius around vertices of a network, naturally leading to a hierarchy of symmetries, in which we progressively use more information and require stronger conditions to affirm that two vertices are symmetric. For the case of 1-local symmetry, we determined regimes of asymptotic local symmetry and asymptotic local asymmetry for an Erdős-Rényi random graph, which indicates the existence of a phase transition, from local symmetry to local asymmetry, between average degrees $n^{1/3}$ and $n^{1/2}$. These findings were complemented by a series of experimental results obtained via simulation.

Several questions naturally follow from our findings. A natural first step would be to determine precisely the average degree in which this phase transition occurs, that is, to close the gap between asymptotic local symmetry and asymmetry regimes. We believe our proof for the local symmetry regimes has potential to be modified and yield a stronger result, by the following reasoning. Recall that, in our proof, we argued that, if a $G(n, p)$ random graph has average degree $o(n^{1/3})$, then with high probability almost all of the neighborhoods of its vertices satisfy the simple structures of a star. The number of such vertices (approximately n) is much larger than the number of distinct structures available, which is approximately $c \cdot np$ for some constant c, due to the degrees in $G(n, p)$ assuming, at most, $c \cdot np$ distinct values. The restriction of average degree $o(n^{1/3})$ is required for the first part of

the argument, however, accepting slightly denser structures (such as stars with a few additional peripheral edges) will allow us to weaken this restriction, with minor adaptation to the rest of the proof. As for the local asymmetry regime, its bounds are fundamentally defined by the requirements of our degree sequence approximation scheme, which makes them hard to extend without resorting to alternative tools.

A more interesting question is whether k-local symmetry also presents this behavior of phase transition for other values of k and, if so, how the location of this phase transition depends on k. For this matter, we believe our proofs for both local symmetry and local asymmetry in the $G(n, p)$ model can be extended. For our proof of local symmetry, we can extend this result by decomposing a neighborhood into "rings" of fixed distance from the center, with the first ring containing vertices at distance exactly 1, and so on — see Figure 7.1 for an illustration. Recall that, in our proof, we used the number of triangles in a graph G to count how many peripheral edges are seen in all 1-neighborhoods of G. In this extension, such peripheral edges correspond to edges inside the first ring of all neighborhoods. A similar reasoning can, for example, relate the number of 5-cycles to the number of edges inside the second rings of all neighborhoods, with proper care to account for overlapping cycles.

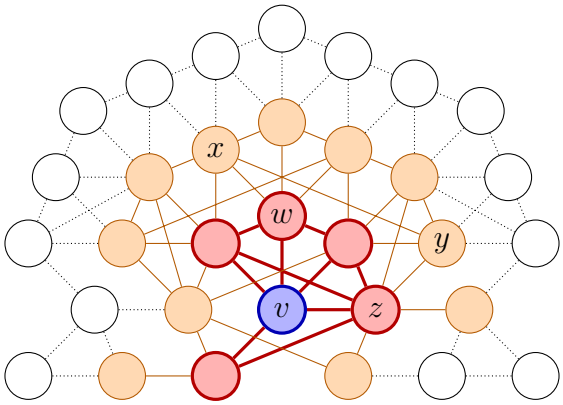


Figure 7.1: Decomposing $\mathcal{N}[v]$ into "rings". Red nodes (at distance 1 from v) form the first ring, orange nodes (distance 2) form the second ring, and white nodes (distance 3) the third ring. A 5-cycle such as $vwxyz$, seen as in $\mathcal{N}[v]$, has two disjoint paths from v to the second ring of $\mathcal{N}[v]$ (vwx and vzy) an a cycle-closing edge inside the second ring (xy).

As for the proof of the asymmetry regime, several facts can be used to extend the result to asymmetry of higher orders. First of all, in the k-neighborhood of any vertex of a $G(n, p)$ random graph, each individual ring is itself a random-sized $G(n, p)$ random graph. Any isomorphism between the neighborhoods of two vertices must perfectly map their respective rings and, therefore, require these rings to be isomorphic themselves. Furthermore, rings from the same neighborhood constitute independent random graphs, due to the disjointness of their vertex sets and to the hierarchical nature of the definition of each ring.

One might also ask whether, and to which extent, real-world networks exhibit local symmetry. This would certainly depend on both the nature of the network and its formation process. However, for certain classes of networks, existing literature allows us to develop some intuition about what answer to expect. For instance, in the social network literature, a number of recent works have attempted to explore the limits of network anonymization [42] [50]. In particular, the *percolation graph matching* (PGM) technique [51] has been successfully used to match common nodes in the Twitter and Flickr networks [50], thus allowing knowledge of one network to be used to break the anonymity of the other. Intuitively, this suggests that each node in these social networks can be uniquely identified structurally within them, from which we would conclude that the networks are globally asymmetric. However, the fact that the PGM technique works by exploiting local neighborhoods would also indicate that these networks exhibit some kind of local asymmetry. Further investigation of this matter would allow us to start exploring the applicability of the local symmetry concept to real-world networks.

It should be noted that, unlike the $G(n, p)$ random graphs, real-world networks exhibit much higher structural diversity of vertices. For instance, networks such as the Internet [66], the Web [67], and scientific collaboration networks [68] are believed to have heavy-tailed degree distributions. This diversity can lead to a similarly diverse behavior regarding global and local symmetry. It is known that most real-world networks are globally symmetric, but the automorphism group of these networks is due to a large number of small subgraphs which are themselves symmetric and comprise vertices with small degrees [69], with high degree vertices being asymmetric to any other vertex in the network. While a simple metric, such as the fraction of vertices with at least one globally or locally symmetric counterpart, would begin to shed a light on this phenomenon, more fine-grained symmetry metrics are desired to capture it in more detail.

The second part of this thesis considered the problem of multiple graph matching, a generalization of the graph matching problem which consists on unveiling a hidden matching between the vertex sets of k similarly-structured graphs. We proposed a mathematical framework for studying this problem, including the concept of hyperpermutation as the representation of a matching across k copies of a set, the $G_k(n, p, s)$ model for multiple correlated random networks, and three distinct mismatch statistics over multiple graphs, namely, full edge mismatch, neutral set and consistency of horizontal mappings. We obtained some preliminary results for the behavior of these statistics in the $G_k(n, p, s)$, including the maximum likelihood estimator for the correct hyperpermutation, and identified key combinatorial difficulties in the mathematical handling of this problem.

It is hard to precise which direction of development is expected to most easily

bear fruit, as the success of any endeavour will be strongly dependent on the identification or proposal of new mathematical tools to tackle this problem. One such tool which we have identified and believe to warrant further investigation is the usage of generating functions in understanding the distribution of mismatch statistics. This technique has been used by Cullina and Kiyavash [55] for the edge mismatch statistic in the graph matching problem on two graphs, resulting in a strong improvement of feasibility results. This leads us to believe that it might be a suitable technique to apply on the full edge mismatch statistic Δ_F, which most directly generalizes the edge mismatch statistic Δ that we had presented beforehand and has been used in this work. However, it is important to point out that using such technique should only help improve how we leverage the correlation of terms in the expression of Δ_F. The hardest challenge we have found in this generalization was how to group hyperpermutation by similarity to the identity. To the best of our knowledge, the corresponding step for two graphs, which is the grouping of permutations, has only been done by taking subsets of permutations with the same order, and this was also the case in the work of Cullina and Kiyavash.

Referências Bibliográficas

[1] "Science Special Issue: Complex Systems and Networks". 1999. Disponível em: <www.sciencemag.org/content/vol325/issue5939/index.dtl>.

[2] FERREIRA, A. A., GONÇALVES, M. A., LAENDER, A. H. "A Brief Survey of Automatic Methods for Author Name Disambiguation", *SIGMOD Rec.*, v. 41, n. 2, pp. 15–26, ago. 2012. ISSN: 0163-5808. doi: 10.1145/2350036.2350040. Disponível em: <http://doi.acm.org/10.1145/2350036.2350040>.

[3] BACKSTROM, L., DWORK, C., KLEINBERG, J. "Wherefore Art Thou R3579X?: Anonymized Social Networks, Hidden Patterns, and Structural Steganography", *Commun. ACM*, v. 54, n. 12, pp. 133–141, dez. 2011. ISSN: 0001-0782. doi: 10.1145/2043174.2043199. Disponível em: <http://doi.acm.org/10.1145/2043174.2043199>.

[4] WEST, D. B. *Introduction to Graph Theory*. 2 ed. , Prentice Hall, September 2000. ISBN: 0130144002.

[5] DALY, E. M., HAAHR, M. "Social Network Analysis for Routing in Disconnected Delay-tolerant MANETs". In: *Proceedings of the 8th ACM International Symposium on Mobile Ad Hoc Networking and Computing*, MobiHoc '07, pp. 32–40, New York, NY, USA, 2007. ACM. ISBN: 978-1-59593-684-4. doi: 10.1145/1288107.1288113. Disponível em: <http://doi.acm.org/10.1145/1288107.1288113>.

[6] BORGATTI, S. P., FOSTER, P. C. "The Network Paradigm in Organizational Research: A Review and Typology", *Journal of Management*, v. 29, n. 6, pp. 991–1013, 2003. doi: 10.1016/S0149-2063_03_00087-4. Disponível em: <http://jom.sagepub.com/conten/29/6/991.abstract>.

[7] BARABASI, A.-L., ALBERT, R. "Emergence of Scaling in Random Networks", *Science*, v. 286, n. 5439, pp. 509–512, 1999. doi: 10.1126/science.286.5439.509. Disponível em: <dx.doi.org/10.1126/science.286.5439.509>.

[8] WATTS, D., STROGATZ, S. "Collective dynamics of 'small-world' networks", *Nature*, v. 393, n. 6684, pp. 409–10, 1998. Disponível em: <dx.doi.org/10.1038/30918>.

[9] ERDÖS, P., RÉNYI, A. "On Random Graphs I", *Publicationes Mathematicae Debrecen*, v. 6, pp. 290, 1959.

[10] GILBERT, E. N. "Random Graphs", *The Annals of Mathematical Statistics*, v. 30, n. 4, pp. 1141–1144, 12 1959. doi: 10.1214/aoms/1177706098. Disponível em: <http://dx.doi.org/10.1214/aoms/1177706098>.

[11] BOLLOBÁS, B. *Random Graphs*. Second ed. , Cambridge University Press, 2001. ISBN: 9780511814068. Disponível em: <http://dx.doi.org/10.1017/CBO9780511814068>. Cambridge Books Online.

[12] ALON, N., SPENCER, J. H. *The Probabilistic Method*. New York, Wiley, 1992.

[13] RÁTH, B. "Mean Field Frozen Percolation", *Journal of Statistical Physics*, v. 137, n. 3, pp. 459–499, 2009. ISSN: 0022-4715. doi: 10.1007/s10955-009-9863-5. Disponível em: <http://dx.doi.org/10.1007/s10955-009-9863-5>.

[14] MATHON, R. "A note on the graph isomorphism counting problem", *Information Processing Letters*, v. 8, n. 3, pp. 131 – 136, 1979. ISSN: 0020-0190. doi: http://dx.doi.org/10.1016/0020-0190(79)90004-8. Disponível em: <http://www.sciencedirect.com/science/article/pii/0020019079900048>.

[15] GAREY, M. R., JOHNSON, D. S. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. New York, NY, USA, W. H. Freeman & Co., 1979. ISBN: 0716710447.

[16] BOOTH, K. S., COLBOURN, C. J. *Problems polynomially equivalent to graph isomorphism*. Computer Science Department, Univ., 1979.

[17] LUKS, E. M. "Isomorphism of graphs of bounded valence can be tested in polynomial time", *Journal of Computer and System Sciences*, v. 25, n. 1, pp. 42 – 65, 1982. ISSN: 0022-0000. doi: http://dx.doi.org/10.1016/0022-0000(82)90009-5. Disponível em: <http://www.sciencedirect.com/science/article/pii/0022000082900095>.

[18] ZEMLYACHENKO, V. N., KORNEENKO, N. M., TYSHKEVICH, R. I. "Graph isomorphism problem", *Journal of Soviet Mathematics*, v. 29,

n. 4, pp. 1426–1481, 1985. ISSN: 1573-8795. doi: 10.1007/BF02104746. Disponível em: <http://dx.doi.org/10.1007/BF02104746>.

[19] BABAI, L. "Graph Isomorphism in Quasipolynomial Time". 2015.

[20] ULLMANN, J. R. "An Algorithm for Subgraph Isomorphism", *J. ACM*, v. 23, n. 1, pp. 31–42, jan. 1976. ISSN: 0004-5411. doi: 10.1145/321921.321925. Disponível em: <http://doi.acm.org/10.1145/321921.321925>.

[21] SCHMIDT, D. C., DRUFFEL, L. E. "A Fast Backtracking Algorithm to Test Directed Graphs for Isomorphism Using Distance Matrices", *J. ACM*, v. 23, n. 3, pp. 433–445, jul. 1976. ISSN: 0004-5411. doi: 10.1145/321958.321963. Disponível em: <http://doi.acm.org/10.1145/321958.321963>.

[22] MCKAY, B. D., PIPERNO, A. "Practical graph isomorphism, {II}", *Journal of Symbolic Computation*, v. 60, pp. 94 – 112, 2014. ISSN: 0747-7171. doi: http://dx.doi.org/10.1016/j.jsc.2013.09.003. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0747717113001193>.

[23] GORI, M., MAGGINI, M., SARTI, L. "Exact and Approximate Graph Matching Using Random Walks", *IEEE Trans. Pattern Anal. Mach. Intell.*, v. 27, n. 7, pp. 1100–1111, jul. 2005. ISSN: 0162-8828. doi: 10.1109/TPAMI.2005.138. Disponível em: <http://dx.doi.org/10.1109/TPAMI.2005.138>.

[24] FOGGIA, P., SANSONE, C., VENTO, M. "A performance comparison of five algorithms for graph isomorphism". In: *in Proceedings of the 3rd IAPR TC-15 Workshop on Graph-based Representations in Pattern Recognition*, pp. 188–199, 2001.

[25] BIGGS, N. *Algebraic Graph Theory*. 2nd ed. , Cambridge University Press, 1993.

[26] ONSAGER, L. "Crystal Statistics. I. A Two-Dimensional Model with an Order-Disorder Transition", *Phys. Rev.*, v. 65, pp. 117–149, Feb 1944. doi: 10.1103/PhysRev.65.117. Disponível em: <http://link.aps.org/doi/10.1103/PhysRev.65.117>.

[27] PEDARSANI, P., GROSSGLAUSER, M. "On the privacy of anonymized networks". In: *ACM KDD*, pp. 1235–1243, 2011.

[28] BAI XIAO, JIAN CHENG, E. R. H. *Graph-Based Methods in Computer Vision: Developments and Applications: Developments and Applications*. IGI Global Research Collection. Information Science Reference, 2012. ISBN: 9781466618923. Disponível em: <https://books.google.com.br/books?id=V66eBQAAQBAJ>.

[29] ERDŐS, P., RÉNYI, A. "Asymmetric graphs", *Acta Mathematica Academiae Scientiarum Hungarica*, v. 14, pp. 295–315, 1963. ISSN: 0001-5954. doi: 10.1007/BF01895716. Disponível em: <http://dx.doi.org/10.1007/BF01895716>.

[30] HOLTON, D. A., SHEEHAN, J. *The Petersen Graph*. Cambridge University Press, 1993.

[31] KIM, J. H., SUDAKOV, B., VU, V. H. "On the Asymmetry of Random Regular Graphs and Random Graphs", *Random Structures and Algorithms*, v. 21, n. 3-4, pp. 216–224, out. 2002. ISSN: 1042-9832. doi: 10.1002/rsa.10054. Disponível em: <http://dx.doi.org/10.1002/rsa.10054>.

[32] GROSSGLAUSER, M., THIRAN, P. "Networks out of Control: Models and Methods for Random Networks". 2014. Disponível em: <http://icawww1.epfl.ch/class-nooc/nooc2014.pdf>.

[33] CHEN, H.-L., LU, H.-I., YEN, H.-C. "On Maximum Symmetric Subgraphs". In: Marks, J. (Ed.), *Graph Drawing*, v. 1984, *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, pp. 372–383, 2001. ISBN: 978-3-540-41554-1. doi: 10.1007/3-540-44541-2_35. Disponível em: <http://dx.doi.org/10.1007/3-540-44541-2_35>.

[34] NAVARRO, G. "A Guided Tour to Approximate String Matching", *ACM Comput. Surv.*, v. 33, n. 1, pp. 31–88, mar. 2001. ISSN: 0360-0300. doi: 10.1145/375360.375365. Disponível em: <http://doi.acm.org/10.1145/375360.375365>.

[35] MCKAY, B. D., WORMALD, N. C. "The degree sequence of a random graph. I. The models", *Random Structures and Algorithms*, v. 11, pp. 97–117, set. 1997. ISSN: 1042-9832. doi: 10.1002/(SICI)1098-2418(199709)11:2⟨97::AID-RSA1⟩3.0.CO;2-O. Disponível em: <http://dl.acm.org/citation.cfm?id=261412>.

[36] KOSTOCHKA, A. V., WEST, D. B. "Chvátal's condition cannot hold for both a graph and its complement", *Discussiones Mathematicae Graph Theory*, v. 26, n. 1, pp. 73–76, 2006.

[37] SKERMAN, F. *Degree Sequences of Random Bipartite Graphs*. Tese de Doutorado, The Australian National University, 2010.

[38] KLEIN, P., YOUNG, N. E. "On the Number of Iterations for Dantzig-Wolfe Optimization and Packing-Covering Approximation Algorithms", *SIAM Journal on Computing*, v. 44, n. 4, pp. 1154–1172, 2015. doi: 10.1137/12087222X. Journal version of [1999]. Published online Aug. 2015.

[39] FINUCAN, H. M. "The Mode of a Multinomial Distribution", *Biometrika*, v. 51, n. 3/4, pp. pp. 513–517, 1964. ISSN: 00063444. Disponível em: <http://www.jstor.org/stable/2334165>.

[40] SIEK, J. G. "An Implementation of Graph Isomorphism Testing". 2001. Disponível em: <http://www.boost.org/libs/graph/doc/isomorphism-impl.pdf>.

[41] BABAI, L., LUKS, E. M. "Canonical Labeling of Graphs". In: *Proceedings of the Fifteenth Annual ACM Symposium on Theory of Computing*, STOC '83, pp. 171–183, New York, NY, USA, 1983. ACM. ISBN: 0-89791-099-0. doi: 10.1145/800061.808746. Disponível em: <http://doi.acm.org/10.1145/800061.808746>.

[42] HAY, M., MIKLAU, G., JENSEN, D., et al. "Resisting structural re-identification in anonymized social networks", *The VLDB Journal*, v. 19, n. 6, pp. 797–823, 2010. ISSN: 0949-877X. doi: 10.1007/s00778-010-0210-x. Disponível em: <http://dx.doi.org/10.1007/s00778-010-0210-x>.

[43] ZASLAVSKIY, M., BACH, F., VERT, J.-P. "Global alignment of protein–protein interaction networks by graph matching methods", *Bioinformatics*, v. 25, n. 12, pp. i259–1267, 2009. doi: 10.1093/bioinformatics/btp196. Disponível em: <http://bioinformatics.oxfordjournals.org/content/25/12/i259.abstract>.

[44] KLAU, G. W. "A new graph-based method for pairwise global network alignment", *BMC Bioinformatics*, v. 10, n. 1, pp. 1–9, 2009. ISSN: 1471-2105. doi: 10.1186/1471-2105-10-S1-S59. Disponível em: <http://dx.doi.org/10.1186/1471-2105-10-S1-S59>.

[45] HE, L., HAN, C. Y., EVERDING, B., et al. "Graph matching for object recognition and recovery", *Pattern Recognition*, v. 37, n. 7, pp. 1557 – 1560, 2004. ISSN: 0031-3203. doi: http://dx.doi.org/10.1016/j.patcog.2003.

12.011. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0031320304000135>.

[46] PAVEL, S., EUZENAT, J. "Ontology Matching: State of the Art and Future Challenges", *IEEE Trans. on Knowl. and Data Eng.*, v. 25, n. 1, pp. 158–176, jan. 2013. ISSN: 1041-4347. doi: 10.1109/TKDE.2011.253. Disponível em: <http://dx.doi.org/10.1109/TKDE.2011.253>.

[47] DOSHI, P., KOLLI, R., THOMAS, C. "Inexact matching of ontology graphs using expectation-maximization", *Web Semantics: Science, Services and Agents on the World Wide Web*, v. 7, n. 2, pp. 90–106, 2009.

[48] JI, S., LI, W., MITTAL, P., et al. "SecGraph: A Uniform and Open-source Evaluation System for Graph Data Anonymization and De-anonymization". In: *24th USENIX Security Symposium (USENIX Security 15)*, pp. 303–318, Washington, D.C., ago. 2015. USENIX Association. ISBN: 978-1-931971-232. Disponível em: <https://www.usenix.org/conference/usenixsecurity15/technical-sessions/presentation/ji>.

[49] SINGH, R., XU, J., BERGER, B. "Global alignment of multiple protein interaction networks with application to functional orthology detection", *Proceedings of the National Academy of Sciences*, v. 105, n. 35, pp. 12763–12768, 2008. doi: 10.1073/pnas.0806627105. Disponível em: <http://www.pnas.org/content/105/35/12763.abstract>.

[50] NARAYANAN, A., SHMATIKOV, V. "De-anonymizing Social Networks". In: *Proceedings of the 2009 30th IEEE Symposium on Security and Privacy*, SP '09, pp. 173–187, Washington, DC, USA, 2009. IEEE Computer Society. ISBN: 978-0-7695-3633-0. doi: 10.1109/SP.2009.22. Disponível em: <http://dx.doi.org/10.1109/SP.2009.22>.

[51] YARTSEVA, L., GROSSGLAUSER, M. "On the Performance of Percolation Graph Matching". In: *Proceedings of the First ACM Conference on Online Social Networks*, COSN '13, pp. 119–130, New York, NY, USA, 2013. ACM. ISBN: 978-1-4503-2084-9. doi: 10.1145/2512938.2512952. Disponível em: <http://doi.acm.org/10.1145/2512938.2512952>.

[52] KORULA, N., LATTANZI, S. "An Efficient Reconciliation Algorithm for Social Networks", *Proc. VLDB Endow.*, v. 7, n. 5, pp. 377–388, jan. 2014. ISSN: 2150-8097. doi: 10.14778/2732269.2732274. Disponível em: <http://dx.doi.org/10.14778/2732269.2732274>.

[53] KAZEMI, E., HASSANI, S. H., GROSSGLAUSER, M. "Growing a Graph Matching from a Handful of Seeds", *Proc. VLDB Endow.*, v. 8, n. 10, pp. 1010–1021, jun. 2015. ISSN: 2150-8097. doi: 10.14778/2794367.2794371. Disponível em: <http://dx.doi.org/10.14778/2794367.2794371>.

[54] PEDARSANI, P., FIGUEIREDO, D. R., GROSSGLAUSER, M. "A bayesian method for matching two similar graphs without seeds". In: *Communication, Control, and Computing (Allerton), 2013 51st Annual Allerton Conference on*, pp. 1598–1607. IEEE, 2013.

[55] CULLINA, D., KIYAVASH, N. "Improved Achievability and Converse Bounds for Erdos-Renyi Graph Matching", *SIGMETRICS Perform. Eval. Rev.*, v. 44, n. 1, pp. 63–72, jun. 2016. ISSN: 0163-5999. doi: 10.1145/2964791.2901460. Disponível em: <http://doi.acm.org/10.1145/2964791.2901460>.

[56] CHIASSERINI, C.-F., GARETTO, M., LEONARDI, E. "De-anonymizing scale-free social networks by percolation graph matching". In: *2015 IEEE Conference on Computer Communications (INFOCOM)*, pp. 1571–1579, April 2015. doi: 10.1109/INFOCOM.2015.7218536.

[57] KAZEMI, E., YARTSEVA, L., GROSSGLAUSER, M. "When can two unlabeled networks be aligned under partial overlap?" In: *53rd Annual Allerton Conference on Communication, Control, and Computing, Allerton 2015, Allerton Park & Retreat Center, Monticello, IL, USA, September 29 - October 2, 2015*, pp. 33–42, 2015. doi: 10.1109/ALLERTON.2015.7446983. Disponível em: <http://dx.doi.org/10.1109/ALLERTON.2015.7446983>.

[58] ZHANG, Y., TANG, J., YANG, Z., et al. "COSNET: Connecting Heterogeneous Social Networks with Local and Global Consistency". In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pp. 1485–1494, New York, NY, USA, 2015. ACM. ISBN: 978-1-4503-3664-2. doi: 10.1145/2783258.2783268. Disponível em: <http://doi.acm.org/10.1145/2783258.2783268>.

[59] HU, J., KEHR, B., REINERT, K. "NetCoffee: a fast and accurate global alignment approach to identify functionally conserved proteins in multiple networks", *Bioinformatics*, p. btt715, 2013.

[60] LIAO, C.-S., LU, K., BAYM, M., et al. "IsoRankN: spectral methods for global alignment of multiple protein networks", *Bioinformatics*, v. 25, n. 12,

pp. i253–i258, 2009. doi: 10.1093/bioinformatics/btp203. Disponível em: <http://bioinformatics.oxfordjournals.org/content/25/12/i253.abstract>.

[61] WILLIAMS, M. L., WILSON, R. C., HANCOCK, E. R. "Multiple Graph Matching with Bayesian Inference", *Pattern Recognition Letters*, v. 18, pp. 080, 1997.

[62] ZHOU, F., DE LA TORRE, F. "Factorized graph matching". In: *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 127–134. IEEE, 2012.

[63] JANSON, S. "Large Deviations for Sums of Partly Dependent Random Variables", *Random Struct. Algorithms*, v. 24, n. 3, pp. 234–248, maio 2004. ISSN: 1042-9832. doi: 10.1002/rsa.v24:3. Disponível em: <http://dx.doi.org/10.1002/rsa.v24:3>.

[64] PEMMARAJU, S. V. "Equitable Colorings Extend Chernoff-Hoeffding Bounds". In: *Proceedings of the Twelfth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '01, pp. 924–925, Philadelphia, PA, USA, 2001. Society for Industrial and Applied Mathematics. ISBN: 0-89871-490-7. Disponível em: <http://dl.acm.org/citation.cfm?id=365411.365811>.

[65] JANSON, S., RUCIŃSKI, A. "The infamous upper tail", *Random Structures & Algorithms*, v. 20, n. 3, pp. 317–342, 2002. ISSN: 1098-2418. doi: 10.1002/rsa.10031. Disponível em: <http://dx.doi.org/10.1002/rsa.10031>.

[66] FALOUTSOS, M., FALOUTSOS, P., FALOUTSOS, C. "On Power-law Relationships of the Internet Topology", *SIGCOMM Comput. Commun. Rev.*, v. 29, n. 4, pp. 251–262, ago. 1999. ISSN: 0146-4833. doi: 10.1145/316194.316229. Disponível em: <http://doi.acm.org/10.1145/316194.316229>.

[67] BRODER, A., KUMAR, R., MAGHOUL, F., et al. "Graph Structure in the Web", *Comput. Netw.*, v. 33, n. 1-6, pp. 309–320, jun. 2000. ISSN: 1389-1286. doi: 10.1016/S1389-1286(00)00083-9. Disponível em: <http://dx.doi.org/10.1016/S1389-1286(00)00083-9>.

[68] NEWMAN, M. E. J. "The Structure of Scientific Collaboration Networks", *Proceedings of the National Academy of Sciences of the United States of America*, v. 98, n. 2, pp. 404–409, 2001. ISSN: 00278424. Disponível em: <http://www.jstor.org/stable/3054694>.

[69] MACARTHUR, B. D., SÁNCHEZ-GARCÍA, R. J., ANDERSON, J. W. "Symmetry in complex networks", *Discrete Applied Mathematics*, v. 156, n. 18, pp. 3525 – 3531, 2008. ISSN: 0166-218X. doi: http://dx.doi.org/10.1016/j.dam.2008.04.008. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0166218X08001881>.