Inférence et réseaux complexes

Thèse

Jean-Gabriel Young

Doctorat en physique
Philosophiæ doctor (Ph. D.)

Québec, Canada

Inférence et réseaux complexes

Thèse

Jean-Gabriel Young

Sous la direction de:

Louis J. Dubé, directeur de recherche
Patrick Desrosiers, codirecteur de recherche

Résumé

Les objets d'études de la science moderne sont souvent complexes : sociétés, pandémies, grilles électriques, niches écologiques, etc. La science des réseaux cherche à mieux comprendre ces systèmes en examinant leur *structure*. Elle fait abstraction du détail, en réduisant tout système à une simple collection de noeuds (les éléments constitutifs du système) connectés par des liens (interactions). Fort d'une vingtaine d'années de recherche, on peut constater que cette approche a mené à de grands succès scientifiques.

Cette thèse est consacrée à l'intersection entre la science des réseaux et l'inférence statistique. On y traite de deux problèmes d'inférence classiques : estimation et test d'hypothèses.

La partie principale de la thèse est dédiée à l'estimation. Dans un premier temps, on étudie un modèle génératif bien connu (le modèle stochastique par blocs), développé dans le but d'identifier les régularités de la structure des réseaux complexes. Les contributions originales de cette partie sont (a) l'unification de la grande majorité des méthodes de détection de régularités sous l'égide du modèle par blocs, et (b) une analyse en taille finie de la cohérence de ce modèle. La combinaison de ces analyses place l'ensemble des méthodes de détection de régularités sur des bases statistiques solides. Dans un deuxième temps, on se penche sur le problème de la reconstruction du passé d'un réseau, à partir d'une seule observation. À nouveau, on l'aborde à l'aide de modèles génératifs, le transformant ainsi en un problème d'estimation. Les résultats principaux de cette partie sont des méthodes algorithmiques permettant de solutionner la reconstruction efficacement, et l'identification d'une transition de phase dans la qualité de la reconstruction, lorsque le niveau d'inégalité des réseaux étudiés est varié.

On se penche finalement sur un traitement par test d'hypothèses des systèmes complexes. Cette partie, plus succincte, est présentée dans un langage mathématique plus général que celui des réseaux, soit celui des complexes simpliciaux. On obtient un modèle aléatoire pour complexe simplicial, ainsi qu'un algorithme d'échantillonnage efficace pour ce modèle. On termine en montrant qu'on peut utiliser ces outils pour tester des hypothèses sur la structure des systèmes complexes réels, via une propriété inaccessible dans la représentation réseau (les groupes d'homologie des complexes).

Abstract

Modern science is often concerned with complex objects of inquiry: intricate webs of social interactions, pandemics, power grids, ecological niches under climatological pressure, etc. When the goal is to gain insights into the function and mechanism of these complex systems, a possible approach is to map their structure using a collection of nodes (the parts of the systems) connected by edges (their interactions). The resulting *complex networks* capture the structural essence of these systems. Years of successes show that the network abstraction often suffices to understand a plethora of complex phenomena.

It can be argued that a principled and rigorous approach to data analysis is chief among the challenges faced by network science today. With this in mind, the goal of this thesis is to tackle a number of important problems at the intersection of network science and statistical inference, of two types: The problems of estimations and the testing of hypotheses.

Most of the thesis is devoted to estimation problems. We begin with a thorough analysis of a well-known generative model (the stochastic block model), introduced 40 years ago to identify patterns and regularities in the structure of real networks. The main original contributions of this part are (a) the unification of the majority of known regularity detection methods under the stochastic block model, and (b) a thorough characterization of its consistency in the finite-size regime. Together, these two contributions put regularity detection methods on firmer statistical foundations. We then turn to a completely different estimation problem: The reconstruction of the past of complex networks, from a single snapshot. The unifying theme is our statistical treatment of this problem, again based on generative modeling. Our major results are: the inference framework itself; an efficient history reconstruction method; and the discovery of a phase transition in the recoverability of history, driven by inequalities (the more unequal, the harder the reconstruction problem).

We conclude with a short section, where we investigate hypothesis testing in complex systems. This epilogue is framed in the broader mathematical context of simplicial complexes, a natural generalization of complex networks. We obtain a random model for these objects, and the associated efficient sampling algorithm. We finish by showing how these tools can be used to test hypotheses about the structure of real systems, using their homology groups.

Table des matières

Liste des figures

Liste des contributions

Les articles qui suivent ont été complétés durant le doctorat. Certains figurent directement dans la thèse (voir avant-propos pour une liste détaillée), alors que plusieurs n'ont pas été reproduits ici, car ne faisant pas partie de la trame narrative de la thèse. Ils sont listés en ordre inverse de parution.

◇ "*Universality of the stochastic block model*" [251]
J.-G. Young, G. St-Onge, P. Desrosiers, and L. J. Dubé
arXiv:1806.04214 (2018)

◇ "*Network archaeology: phase transition in the recoverability of network history*" [249]
J.-G. Young, L. Hébert-Dufresne, E. Laurence, C. Murphy, G. St-Onge and P. Desrosiers
arXiv:1803.09191 (2018)

◇ "*Exact analytical solution of irreversible binary dynamics on networks*" [134]
E. Laurence, **J.-G. Young**, S. Melnik and L. J. Dubé
Phys. Rev. E, **97**, 032302 (2018)

◇ "*Phase transition of the susceptible-infected-susceptible dynamics on time-varying configuration model networks*" [225]
G. St-Onge, **J.-G. Young**, E. Laurence, C. Murphy, L. J. Dubé
Phys. Rev. E **97**, 022305 (2018)

◇ "*Susceptible-infected-susceptible dynamics on the rewired configuration model*" [226]
G. St-Onge, **J.-G. Young**, E. Laurence, C. Murphy, L. J. Dubé
arXiv:1701.01740 (2017)

◇ "*Construction of and efficient sampling from the simplicial configuration model*" [250]
J.-G. Young, G. Petri, F. Vaccarino, and A. Patania
Phys. Rev. E, *96*, 032312 (2017)

◇ "*Strategic tradeoffs in competitor dynamics on adaptive networks*" [99]
L. Hébert-Dufresne, A. Allard, P.-A. Noël, **J.-G. Young**, and E. Libby
Sci. Rep., *7*, 7576 (2017)

◇ "*Finite size analysis of the detectability limit of the stochastic block model*" [247]
J.-G. Young, P. Desrosiers, L. Hébert-Dufresne, E. Laurence and L. J. Dubé
Phys. Rev. E, *95*, 062304 (2017)

◇ "*Growing networks of overlapping communities with internal structure*" [248]
J.-G. Young, L. Hébert-Dufresne, A. Allard and L. J. Dubé
Phys. Rev. E **94**, 022317 (2016)

◇ "*Constrained growth of complex scale-independent systems*" [100]
L. Hébert-Dufresne, A. Allard, **J.-G. Young** and L. J. Dubé
Phys. Rev. E **93**, 032304 (2016)

◇ "*Complex networks as an emerging property of hierarchical preferential attachment*" [102]
L. Hébert-Dufresne, E. Laurence, A. Allard, **J.-G. Young** and L. J. Dubé
Phys. Rev. E, **92**, 062809 (2015)

◇ "*General and exact approach to percolation on random graphs*" [9]
A. Allard, L. Hébert-Dufresne, **J.-G. Young** and L. J. Dubé
Phys. Rev. E, **92**, 062807 (2015)

◇ "*A shadowing problem in the detection of overlapping communities*" [246]
J.-G. Young, A. Allard, L. Hébert-Dufresne and L. J. Dubé
PLoS ONE **10**, e0140133 (2015)

Liste d'abbréviations et de notations

Abbréviations

CM	Modèle de configuration (*configuration model)*
EM	Algorithme espérance–maximisation (*expectation maximization algotirhm*)
ERGM	Modèle exponentiel de graphes aléatoires (*exponential random graph model*)
GMGM	Modèle modulaire général (*general modular graph model*)
LG	Graphe adjoint (*line graph*)
MAP	Maximum a posteriori
MC	Monte-Carlo
MCMC	Méthode de Monte-Carlo par chaînes de Markov (*Monte Carlo Markov Chain*)
MMO	Maximum du recouvrement moyen (*maximum mean overlap*)
MMSE	Minimum de la moyenne de l'erreur quadratique (*minimum mean-squared error*)
MPE	Extraction de patrons mésoscopiques (*mesoscopic pattern extraction*)
SBM	Modèle stochastique par blocs (*stochastic block model*)
SCM	Modèle de configuration simpliciale (*simplicial configuration model)*
PA	Modèle d'attachement préférentiel (*preferential attachment*)

Asymptotes

$f(n) = o(g(n))$	La fonction $f(n)$ est dominée par $g(n)$ $\lim_{n\to\infty} \left\lvert \frac{f(n)}{g(n)} \right\rvert = 0$
$f(n) = O(g(n))$	La fonction $f(n)$ est du même ordre que $g(n)$ $\exists\, c > 0$ tel que $\lim_{n\to\infty} \left\lvert \frac{f(n)}{g(n)} \right\rvert \leq c$

Symboles

$\langle f(x) \rangle$	Espérance de $f(x)$.
$\hat{f}$	Estimateur de f.
$\boldsymbol{X}$	Matrice d'éléments x_{ij}.
$\boldsymbol{x}$	Vecteur rangée.
$\boldsymbol{x}^{\mathsf{T}}$	Vecteur colonne.
$\mathbf{1}$	Vecteur de 1.
$\lvert y \rvert$	Cardinalité de l'ensemble y.
$\mathbb{I}[X]$	Fonction indicatrice. $\mathbb{I}[x] = 1$ si l'énoncé x est vraie, et $\mathbb{I}[x] = 0$ autrement.
δ_{ij}	Delta de Kronecker. $\delta_{ij} = 1$ si $i = j$, et $\delta_{ij} = 0$ autrement.
$\delta(x)$	Delta de Dirac. Défini par $\int \delta(x)dx = 1$ et $\int \delta(x - x_0) f(x) dx = f(x_0)$.
$G \sim X(\theta)$	La variable G est tirée du modèle ou de la distribution X, de paramètre θ.
$f(x) \sim g(x)$	La fonction $f(x)$ est équivalente à la fonction $g(x)$.
$\prod\limits_{i \leq j}^{n}$	Produit sur les paires (i, j) tel que $1 \leq i \leq j \leq n$, où n est omis s'il n'est pas ambiguë.
$\prod\limits_{i < j}^{n}$	Produit sur les paires (i, j) tel que $1 \leq i < j \leq n$, où n est omis s'il n'est pas ambiguë.
$[n]$	Ensemble d'entiers contigus $\{1, \ldots, n\}$.

Like a tropical storm, I, too,
may one day become
“better organized.”

Lydia Davis
Samuel Johnson is Indignant (2001)

Remerciements

Entreprendre des études supérieures, c'est en quelque sorte choisir de vivre dans l'antichambre du monde, encore un moment. Des succès et inévitables moments creux des dernières années, l'impression qui me reste en est une de grande joie et de gratitude, pour laquelle je suis redevable à de nombreuses personnes. Je leur dédie ces pages.

Mes premiers remerciements vont à Louis J. Dubé, capitaine du navire Dynamica. J'ai passé l'entièreté de ma carrière scientifique sous son égide, de mes premiers pas en tant que stagiaire estivale il y a sept ans, jusqu'à l'achèvement du présent ouvrage. Son expertise de navigateur du monde académique et son ouverture d'esprit sont d'inestimables atouts qui ont rendu l'aventure possible. Je le remercie pour sa confiance, sa patience et pour la liberté unique qu'il accorde aux membres de son groupe. La force de Dynamica est sa cohésion et ses hauts standards, deux qualités ancrées dans un respect et une exigence mutuelle, instiguées par LJD. Cette thèse est le produit de l'environnement intellectuel stimulant qui en résulte.

Patrick Desrosiers s'est joint à la barre du navire en tant que co-directeur pour mon doctorat. Ceux qui l'ont fréquenté reconnaîtront son influence dans mes travaux. Sa connaissance encyclopédique de la science est une source inépuisable d'inspiration, et son amour (contagieux) de la clarté mathématique n'a laissé aucune chance aux raisonnements imprécis. Mes travaux ont grandement bénéficié de son regard critique.

Je tiens également à remercier les professeurs Jean-François Fortin, Pierre Mathieu, Nicolas Doyon et Daniel Larremore d'avoir accepté de siéger sur mon comité d'évaluation. Même si le traitement mathématique des réseaux complexes est proche de celui utilisé en physique statistique, les sujets abordés dans cette thèse s'éloignent quelque peu de la physique théorique plus classique. J'apprécie donc d'autant plus le travail des examinateurs, que j'entraîne ici vers des horizons inconnus.

On dit souvent qu'il faut un village pour élever un enfant ; fort de l'expérience des dernières années, je peux dire sans l'ombre d'un doute que le dicton s'applique également au doctorat. Je suis en particulier tributaire de mes collaborateurs de longue date : Laurent Hébert-Dufresne pour le torrent d'idées, les opportunités et son mentorat ; Antoine Allard pour sa précision et sa constance ; Edward Laurence pour son regard frais et perspicace, qui va à l'essentiel. Tout aussi important sont les collaborateurs plus récents : Guillaume St-Onge et Charles Murphy qui se sont joints à Dynamica il y a deux ans ; Pierre-André Noël que j'ai côtoyé au tout début de mon séjour dans Dynamica et que j'ai eu le plaisir de revoir, dans le cadre de collaborations récentes ; Eric Libby du Santa Fe Institute ; Alice Patania, Giovanni Petri et Francesco Vaccarino de la fondation ISI à Turin. Leurs points de vue variés ont mené à des projets beaucoup plus riches.

Dynamica c'est aussi une grande famille. Je salue donc tout spécialement ceux qui sont passés dans le giron du groupe. Nos échanges autour d'un café, d'une poutine de Ti-Oui ou d'un Kem'bourek de chez Habil ont contribué à en faire des expériences inoubliables. Du côté réseau : Vincent Marceau, ainsi que les nouveaux impétrants Xavier Roy-Pomerleau et Vincent Thibeault. Du côté lumière : Joey Dumont avec qui j'ai entrepris la maîtrise ; Denis Gagnon et Guillaume Painchaud-April qui complétaient déjà leur formation lors de mon arrivée ; Jean-Luc Déziel, l'irréductible gaulois.

Hors du cercle académique immédiat, je tiens d'abord à remercier les amis qui ont suivi un parcours semblable au mien : Ludovic, frère d'arme des premières heures, toujours là pour partager les bons et les mauvais coups, avec le mot juste ; Sophie et Elena pour leurs conseils académiques en direct ; la cohorte du baccalauréat, encore unie après tant d'années.

Je voudrais aussi saluer les amis et la famille, trop nombreux pour mentionner ici : du complexe, d'ALBATROS, des téteux, de la rue des pauvres et du club de lecture. Votre présence permet de remettre tout en perspective et de profiter de ce qui compte réellement.

Finalement, je voudrais remercier tout particulièrement mes proches : mes parents Louise et Ronald, pour les encouragements inconditionnels, l'éveil intellectuel, les attentions ; Sarah et Raphaël pour la fratrie et les rires ; et Alice pour sa joie de vivre contagieuse, l'inspiration et l'écoute. Elle a contribué à ces pages plus qu'elle ne saurait l'imaginer.

Avant-propos

Les articles qui suivent sont directement incorporés à la thèse. Le contenu n'en a été modifié que pour se conformer au format exigé par la Faculté des études supérieures et postdoctorales de l'Université Laval. Dans tous les cas, l'idéation, les développements théoriques, les simulations numériques de même que la rédaction du manuscrit sont principalement les fruits du travail du premier auteur. Les autres auteurs participent à l'élaboration des projets, supportent les travaux théoriques et numériques, de même que révisent et améliorent le manuscrit. La seule exception est le Chapitre 5, qui est le résultat d'une contribution égale avec A. Patania.

◇ CHAPITRE 2
"*Universality of the stochastic block model*" [251]
J.-G. Young, G. St-Onge, P. Desrosiers, and L. J. Dubé
arXiv:1806.04214
Sous presse. *Phys. Rev. E* (soumission : juin 2018, révision : août 2018)

◇ CHAPITRE 3
"*Finite size analysis of the detectability limit of the stochastic block model*" [247]
J.-G. Young, P. Desrosiers, L. Hébert-Dufresne, E. Laurence and L. J. Dubé
Phys. Rev. E, *95*, 062304 (2017)

◇ CHAPITRE 4
"*Network archaeology: phase transition in the recoverability of network history*" [249]
J.-G. Young, L. Hébert-Dufresne, E. Laurence, C. Murphy, G. St-Onge and P. Desrosiers
arXiv:1803.09191
En révision. *Proc. Natl. Acad. Sci. U.S.A* (soumission : mai 2018)

◇ CHAPITRE 5
"*Construction of and efficient sampling from the simplicial configuration model*" [250]
J.-G. Young, G. Petri, F. Vaccarino, and A. Patania
Phys. Rev. E, *96*, 032312 (2017)

Introduction

Trois époques

Warren Weaver, un des pères fondateurs de l'étude de la complexité, divisait la science moderne en trois grandes époques. D'abord, une époque de la science du «simple», caractérisée par des problèmes aisément et précisément quantifiables (e.g., physique Newtonienne). Ensuite, une époque de la science de la complexité désorganisée, où les problèmes sont devenus plus complexes, mais toutefois simples à décrire d'un point de vue *global* (e.g., physique des gaz). Puis finalement, une époque annoncée de la science de la complexité organisée—la nôtre—, caractérisée par des systèmes dont les composantes sont corrélées, et révèlent des comportements globaux *émergents*, surprenants (e.g., éconophysique).

Dans son célèbre essai *Science and Complexity* [237], Weaver soutient que cette nouvelle science (1948) devra solutionner des problèmes fondamentalement plus ardus que ceux qui ont été attaqués auparavant. Des difficultés qui sont d'une part dues aux comportements émergents des systèmes étudiés, et d'autre part dues au fait que les problèmes de la science de la complexité transcendent les frontières disciplinaires établies. Ainsi, poursuit-il, il faudra développer de nouvelles techniques mathématiques et computationnelles, car

> *... problems [of organized complexity] are just too complicated to yield to the old nineteenth century techniques which were so dramatically successful on two-, three-, or four-variable problems of simplicity. These new problems, moreover, cannot be handled with the statistical techniques so effective in describing average behavior in problems of disorganized complexity.*
>
> —W. Weaver (1948)

Complexité moderne, physique statistique et science des réseaux

L'histoire a, en bonne partie, donné raison à Weaver. La complexité est maintenant l'affaire de plusieurs disciplines, et ses méthodes statistiques ont dû évoluer pour se conformer aux réalités des problèmes de la complexité organisée [155].

Une grande partie de ces transformations est provenue de travaux en physique statistique [155, 227]. La deuxième moitié du XXème siècle, en particulier, a été marquée par une foule de découvertes physiques qui ont trouvé des applications hors des confins de la discipline. On pense par exemple aux méthodes d'échantillonnage développées par Metropolis et Hastings (chaîne de Markov Monte Carlo) et John Von Neumann (échantillonnage préférentiel), maintenant populaire en statistique et en apprentissage automatique [11]. Ou encore aux travaux de Edward, Anderson, Sherrington et Kirkpatrick (et plusieurs autres) sur les matériaux magnétiques, qui nous ont permis de mieux comprendre plusieurs phénomènes complexes issus de domaines aussi variés que l'écologie, la médecine ou l'informatique théorique [227].

Depuis peu, la science de la complexité s'attaque maintenant à des problèmes réseaux [168], c'est-à-dire des problèmes où on doit, par-dessus tout, comprendre la structure du système pour arriver à une solution. La physique joue, encore une fois, un rôle important dans ce récent axe de recherche [6], car l'approche de la science des réseaux est une approche universelle : on y fait abstraction du détail et on se concentre sur la structure (interactions, connections), en cherchant à extirper des propriétés universelles. Fort d'une vingtaine d'années de recherche, on peut constater que cette approche à la complexité a mené à de grands succès scientifiques [15]. Notre compréhension des réseaux nous permet déjà, par exemple, de contrôler des systèmes neuronaux [142], de comprendre l'effet des comportements humains sur les épidémies [8, 215], d'identifier des mécanismes menant à l'émergence des réseaux complexes réels [99, 102], ou encore d'identifier les principes organisationnels des réseaux réels [169].

Vers un traitement plus rigoureux des réseaux

Après des années d'exploration fructueuse, la science des réseaux doit maintenant affronter le défi de la consolidation des acquis, afin de mieux préparer sa prochaine lancée. En effet, plusieurs travaux classiques de la science des réseaux reposent, de par leur nature exploratoire, sur des heuristiques ou des méthodes *ad hoc*. Ceci ne veut pas nécessairement dire que leurs conclusions sont fausses, mais plutôt qu'il y a place pour établir des bases plus solides. Ce genre de consolidation a déjà eu lieu pour certains aspects de la science des réseaux ; on peut par exemple penser à la ré-interprétation récente de la détection de communautés [176] comme un problème d'inférence [254]. Le résultat en vaut la peine, puisqu'on comprend maintenant beaucoup mieux le problème de la détection, les méthodes associées, et leurs limitations [158, 252].

Plusieurs ont fait de cette consolidation leur cheval de bataille [110]. On peut même dire que l'établissement de fondements solides pour la science des réseaux est un programme de recherche en émergence. La présente thèse est écrite dans cette mouvance. Son thème unificateur est la recherche de bases rigoureuses pour un certain nombre de problèmes réseaux, ici présentés comme des problèmes d'inférence statistique.

Organisation de la thèse

Le chapitre 1 se veut une introduction pédagogique aux concepts abordés dans la thèse, car celle-ci est principalement présentée sous forme d'articles scientifiques, plus austères. Ainsi, le but de ce premier chapitre est d'expliquer les bases d'une analyse statistique rigoureuse pour réseaux complexes. Pour y arriver, on introduit d'abord plusieurs modèles aléatoires de réseaux complexes, qu'on utilise ensuite pour définir une approche d'inférence statistique par modèles génératifs. On motive cette analyse avec des problèmes pratiques, ce qui nous permet finalement d'aborder la partie algorithmique de l'inférence réseau.

Le coeur de la thèse s'articule en deux grandes parties, suivies d'un épilogue.

Dans la première partie, on traite de l'inférence *structurelle*, au niveau mésoscopique. On y développe et analyse des outils numériques permettant d'identifier les régularités mésoscopiques d'un réseau, c'est à dire des groupes ou blocs de noeuds jouant le même rôle structurel dans le réseau. Ces développements sont motivés par la promesse qu'une description mésoscopique d'un système complexe est plus facile à appréhender, plus utile et même plus informative qu'une description microscopique détaillée [155]. Le premier chapitre de cette partie, le chapitre 2, établit le contexte. On y montre que l'analyse d'une grande classe d'algorithmes d'inférence peut être réduite à l'analyse d'un modèle *universel*, appelé le modèle stochastique par blocs (SBM) [105]. Fort de cette conclusion, on procède à l'analyse de la cohérence du SBM en taille finie (chapitre 3). L'analyse en taille finie nous permet de définir une mesure de la difficulté du problème d'inférence, et d'obtenir des classes d'équivalences pour cette mesure (plus précisément, de déterminer si les régularités cachées de deux réseaux sont aussi difficiles à identifier). En raison de l'universalité du SBM, les résultats de cette première partie—tous originaux—s'appliquent à une large classe d'algorithmes d'inférence. Ils viennent complémenter la littérature grandissante de la théorie moderne du SBM [1].

Dans la deuxième partie, on traite de la structure des systèmes complexes en tant que signature d'un processus temporel. Le but de l'inférence n'est plus d'identifier les régularités du réseau, mais plutôt de reconstruire son histoire (de « *l'archéologie réseau* »). Ce point de vue original permet de prendre en compte les corrélations structurelles ; elles sont typiquement ignorées par les approches mésoscopiques [76, 78]. Bien que les concepts et méthodes développés dans cette partie soient en principe généralisables à plusieurs modèles de croissance, on concentre nos efforts sur un seul modèle, soit une généralisation du célèbre attachement préférentiel [16]. On montre que l'histoire d'un réseau peut être reconstruite—parfois quasi parfaitement—à l'aide de la distribution *a posteriori* sur les trajectoires historiques du réseau. On montre également qu'on peut évaluer cette distribution efficacement à l'aide d'une méthode de réduction de variance. Ces travaux ouvrent la porte à l'inférence mésoscopique à l'aide de modèles de croissance avec structure (voir les références [97, 102, 248]).

Un épilogue suit les deux parties principales de la thèse. On y traite d'un aspect complètement différent de l'inférence statistique pour systèmes complexes ; plutôt que de chercher les propriétés cachées d'un système (estimation de paramètres), on se concentre plutôt sur le test d'hypothèses [36]. L'objectif de l'épilogue est double. Du point de vue conceptuel, on veut étendre la portée de nos outils en s'attaquant à des systèmes dont la structure est mal représentée par les réseaux complexes. Du point de vue appliqué, on cherche à développer une approche aux tests d'hypothèses qui est pratique (numériquement efficace). On atteint ces deux objectifs en introduisant un modèle aléatoire de complexes simpliciaux—une généralisation des réseaux—, ainsi qu'un échantillonneur efficace pour cet ensemble. Ceci nous permet d'établir numériquement les propriétés des complexes simpliciaux non-structurés, et donc de mettre en contexte les propriétés des systèmes réels associés.

Chapitre 1

Théorie et outils de l'inférence sur réseaux complexes

Les sujets abordés dans cette thèse se situent à la croisée de la physique statistique, de l'inférence et de l'informatique théorique [153, 158, 252]. Puisqu'il s'agit d'un ensemble de domaines *a priori* disparates, ce premier chapitre se veut une introduction pédagogique, permettant d'établir le vocabulaire et les outils qui seront utilisés dans l'ensemble de la thèse. La présentation est organisée en trois mouvements qui suivent le même *leitmotiv*. Dans un premier temps, on introduit des modèles aléatoires de réseaux complexes, objets d'étude de la thèse (Sec. 1.1). On les utilise ensuite pour motiver une approche par modèles génératifs de l'inférence statistique réseau (Sec. 1.2), puis on conclut en introduisant des algorithmes permettant d'appliquer les modèles et méthodes développés au cours du chapitre (Sec. 1.3). On terminera ce chapitre introductif en ayant survolé tous les outils nécessaires à un traitement statistique rigoureux des *réseaux complexes*.

1.1 Réseaux complexes

Un réseau complexe est un objet mathématique décrivant la structure d'un système complexe. La *science des réseaux* se penche sur les systèmes pour lesquelles cette structure suffit à déterminer, comprendre, et expliquer les propriétés importantes observées. Le succès incontestable de la science des réseaux est dû au fait qu'il s'agit souvent d'un mode d'analyse suffisamment puissant, qui permet d'expliquer beaucoup avec peu [15, 89]. Précédant toute forme de traitement réseau, il y a d'abord une étape non triviale d'abstraction, lors de laquelle on identifie les éléments constitutifs du système complexe d'intérêt, ainsi que leurs interactions [170]. Cet exercice, plus ou moins évident, mène à une représentation du système, que nous considérons dès lors comme connue, et le point de départ de notre analyse.

1.1.1 Structure

Words are a pretty fuzzy substitute for mathematical equations.
—Isaac Asimov

La structure d'un réseau $G(V,E)$ est spécifiée par un ensemble de noeuds V et un ensemble de liens E, formé de paires de noeuds (v_i, v_j) tirées de V. Le but du processus d'abstraction mentionné ci-haut est de représenter les éléments constitutifs du système par les noeuds $v_i \in V$, et d'encoder leurs interactions à l'aide de liens. Reflétant la nature complexe de ces interactions, les liens peuvent être, en général, munis d'une direction et / ou accompagnés d'un poids. On se concentrera toutefois principalement sur les réseaux dits simples, sans poids, sans direction, et dans lesquels un noeud n'est jamais connecté à lui-même[1]. Ainsi on ne distinguera pas (v_i, v_j) de (v_j, v_i). On définit les concepts suivants.

Matrice d'adjacence. On encode la structure d'un réseau dans une matrice d'adjacence $\boldsymbol{A}$, où l'élément a_{ij} donne le poids de la connexion $v_j \to v_i$. Cette matrice est de dimension $N \times N$, où $N = |V|$ est le nombre de noeuds. Elle compte $2M$ éléments non nuls, où $M = |E|$ est le nombre de liens (la *taille* du réseau). Dans le cas d'un réseau simple, les paires (v_i, v_j) sont non ordonnées et la matrice $\boldsymbol{A}$ est symétrique ; de plus puisque les liens n'ont pas de poids, la matrice est binaire. Finalement, l'absence de boucles implique que la diagonale de la matrice est nulle.

Degré. Le degré k_i d'un noeud v_i correspond à son nombre de voisins. On dit que le noeud v_j est voisin du noeud v_i s'il existe un lien les connectant, i.e., si $(v_i, v_j) \in E$. On peut facilement obtenir le degré d'un noeud à l'aide de la matrice d'adjacence $\boldsymbol{A}$, en sommant ses rangées ou colonnes, ce qui nous donne une équation matricielle pour le vecteur des degrés $\boldsymbol{k}$:

$$k_i = \sum_{j=1}^{N} a_{ij} = \sum_{j=1}^{N} a_{ji} \quad \Leftrightarrow \quad \boldsymbol{k} = \boldsymbol{A}\mathbf{1} = \boldsymbol{A}^\intercal \mathbf{1}\,, \tag{1.1}$$

où $\mathbf{1} = [1,1,\ldots,1]^\intercal$.

Degré moyen et moments. Le degré caractérise la connectivité des noeuds sur une base individuelle. Mais on peut également quantifier la connectivité du réseau entier, en considérant les moments de la séquence $\boldsymbol{k}$. Par exemple, le degré moyen

$$\langle k \rangle = \sum_{i=1}^{N} \frac{k_i}{N} = \frac{1}{N} \mathbf{1}^\intercal \boldsymbol{A} \mathbf{1} \tag{1.2}$$

est un bon indicateur de la connectivité du réseau. Si $\langle k \rangle < 1$, les noeuds sont en moyenne déconnectés, alors que $\langle k \rangle \gg 1$ indique une forte connectivité globale. Les moments supérieurs

[1] Sauf dans les chapitres 4 et 5, où cette exception est clairement identifiée.

$\langle k^\ell \rangle$ et les moments centrés $\langle \tilde{k}^\ell \rangle$, de la forme

$$\langle k^\ell \rangle = \sum_{i=1}^{N} \frac{k_i^\ell}{N} \qquad \text{et} \qquad \langle \tilde{k}^\ell \rangle = \sum_{i=1}^{N} \frac{(k_i - \langle k \rangle)^\ell}{N} \,, \tag{1.3}$$

viennent complémenter cette information, par exemple en quantifiant la concentration des degrés.

Densité. Connaître le degré moyen n'est pas toujours utile, puisqu'il s'agit d'un nombre absolu, sans contexte. On utilise donc aussi la *densité* ρ pour quantifier la connectivité d'un réseau. Elle est donnée par le nombre de liens dans le réseau, divisé par le nombre maximal de liens pouvant être placés entre les N noeuds du réseau. Elle est donc reliée au degré via

$$\rho = \frac{M}{\binom{N}{2}} = \frac{\langle k \rangle}{N} + O(N^{-2}) \,. \tag{1.4}$$

L'intérêt de ρ est qu'on peut l'utiliser pour définir des *régimes de densité*. Ceux-ci permettent de classer un réseau et, par exemple, de prédire si une méthode d'inférence ou un calcul approximatif fonctionnera bien sur le réseau en question[2]. Les régimes de densité sont rigoureusement définis en termes de *séquences* de réseaux, de plus en plus grands. Si une séquence de réseaux a une densité ρ_N tendant vers 0 dans la limite $N \to \infty$, alors on dit ces réseaux 'creux' (*sparse*) ; si la limite est finie, alors les réseaux sont dits 'denses'. La démarcation n'est pas aussi claire dans le cas fini. On dit simplement qu'un réseau est creux si $\rho \approx 0$ et qu'il est dense autrement. Finalement, notons que dans la limite $N \to \infty$, $\rho = 0$ n'implique pas l'absence de liens : il suffit d'avoir $\langle k \rangle = o(N)$ pour que les réseaux soient creux. Ainsi, même une séquence de réseau dont le degré moyen croît avec N peut être considérée creuse.

1.1.2 Modèles aléatoires

> *All models are wrong, but some are useful.*
>
> —George Box

Pour appréhender la structure d'un réseau réel de grande envergure, on commence normalement par calculer les quantités définies ci-haut—et plusieurs autres [170]—, à partir de sa matrice d'adjacence. Ceci nous donne une première impression de la structure d'un objet autrement trop difficile à visualiser. Or, cette première impression est souvent limitée puisqu'un nombre seul a rarement une signification intrinsèque. On doit établir des points de comparaison pour que ces quantités prennent tout leur sens [79, 182].

[2] On verra des exemples de méthodes d'inférence qui fonctionnent uniquement dans le régime $\rho \gg 0$ (Chap. 3), mais aussi des exemples de calculs seulement valides dans le régime $\rho \approx 0$ (e.g. chapitre 5, ainsi que le présent chapitre).

Afin d'illustrer ce concept, considérons par exemple le *coefficient d'agrégation moyen*[3] d'un réseau. Savoir que ce coefficient est égal à une valeur $C(G)$ donnée n'est pas très informatif. Il est beaucoup plus intéressant d'apprendre, par exemple, que la valeur *typique* $\langle C(G)\rangle$ de ce coefficient est en moyenne beaucoup plus forte que pour des réseaux de même densité. Dans ce cas, la comparaison à une valeur typique nous dit que la propriété $C(G)$ est surprenante ; l'agrégation n'est pas seulement expliquée par la densité. Ce genre de comparaison nécessite évidemment de pouvoir calculer et définir les valeurs $\langle C(G)\rangle$ de référence. Adoptant le point de vue de *l'inférence statistique* pour réseaux [36], nous calculerons ces valeurs en termes d'ensembles nuls de réseaux (i.e., des modèles aléatoires). On présente donc quelques modèles aléatoires standard dans cette section.

Notons avant de se lancer qu'il y a deux aspects importants à garder en tête lors de l'élaboration d'un modèle aléatoire : sa définition, mais aussi la question de l'échantillonnage [45]. Dans les modèles dits mécanistes, ces deux aspects sont trivialement reliés, puisqu'on définit directement les modèles mécanistes en termes d'une méthode de construction (voir, e.g., **Attachement préférentiel** ci-bas). Dans les modèles statistiques purs, la séparation est bien réelle et potentiellement problématique[4]. En effet, s'il est facile de postuler une distribution quelconque sur un ensemble de graphes, échantillonner cette distribution correctement et efficacement est beaucoup plus difficile [79]. Ainsi, bien qu'on se concentrera dans un premier temps sur les définitions, on devra également discuter du problème de l'échantillonnage ; on y revient à la fin du chapitre (c.f. Sec 1.3).

Modèle de Erdős-Rényi / Albert.[5] Historiquement le premier modèle de graphe aléatoire, le modèle de *Erdős-Rényi* (ER) est le plus simple qu'on puisse définir. À l'instar des ensembles de la physique statistique, il existe deux versions du modèle, qu'on appelle microcanonique et canonique, respectivement.

Dans la version microcanonique, on place une probabilité $P(G) = 1/|\Omega_{\text{ER}}(N,M)|$ sur tous les graphes de l'ensemble $\Omega_{\text{ER}}(N,M)$ des graphes de N noeuds et M liens, où $0 \leq M \leq \binom{N}{2}$. Ainsi une réalisation de l'ensemble est n'importe quel graphe dans $\Omega_{\text{ER}}(N,M)$, choisi uniformément. Cet ensemble est de taille $\binom{\binom{N}{2}}{M}$, ce qui fixe la probabilité $P(G)$ de tirer un graphe particulier. L'ensemble est dit microcanonique, puisque l'observable—le nombre de liens—est fixée exactement plutôt qu'en moyenne.

[3] Il s'agit de la moyenne des coefficients d'agrégation locaux, donné pour le noeud v_i par le nombre de triangles $\{(v_i,v_j),(v_i,v_k),(v_k,v_j)\}$ centrés sur v_i, divisé par le nombre de triades $\{(v_i,v_j),(v_i,v_k)\}$ incluant v_i, avec $i \neq j \neq k$. Ce coefficient mesure la tendance des voisins de v_i à être interconnectés.

[4] Voir par exemple le chapitre 5, dont l'objectif est de réconcilier une définition et la méthode d'échantillonnage associée.

[5] *Note historique* : Le «modèle de Erdős-Rényi» désigne en fait deux modèles très semblables, développés indépendamment. Le premier est dû à Edgar Gilbert des laboratoires Bell ; il s'agit d'une formulation canonique—au sens de la physique statistique—du concept de graphes aléatoires sans structure [84]. Le deuxième modèle est dû à Paul Erdős et Alfred Rényi [68, 69] ; il s'agit d'une formulation microcanonique du même concept. À cause d'un accident de l'histoire, on attribue les deux modèles à Erdős et à Rényi, bien qu'ils ne soient pas (parfaitement) identiques [27] et qu'ils aient été découverts simultanément.

Dans la version canonique du modèle ER, on place un lien avec probabilité p, entre chaque paire de noeuds (v_i, v_j), de sorte que le graphe simple de matrice d'adjacence $\boldsymbol{A}$ est générée avec probabilité

$$P(G) = \prod_{i<j} \left[(1-p)(1-a_{ij}) + p \, a_{ij} \right] \equiv \prod_{i<j} (1-p)^{1-a_{ij}} p^{a_{ij}} \,, \tag{1.5}$$

où le produit est pris sur chaque paire de noeuds distincte. Les éléments de la matrice $\boldsymbol{A}$ agissent ici comme des variables indicatrices, permettant de sélectionner la probabilité associée à chaque lien / non-lien. On peut donc plus simplement écrire

$$P(G) = (1-p)^{\binom{N}{2}-M} p^M \tag{1.6}$$

où M est le nombre de liens dans G. De cette expression pour $P(G)$, on déduit immédiatement que le nombre de liens dans un graphe généré par le modèle canonique est tiré de la distribution binomiale : on retient chacun des $\binom{N}{2}$ liens possibles avec probabilité p. Ce modèle est dit canonique car le nombre de liens est fixé *en moyenne*, une moyenne qui est contrôlée par le paramètre $p \in [0,1]$ (une température, en quelque sorte).

Puisque M est déterminé par une loi binomiale, un graphe $G \sim \mathrm{ER}(p)$ a en moyenne $\langle M(p) \rangle = \binom{N}{2} p$ liens, une densité p, et ses noeuds ont en moyenne $\langle k \rangle = p(N-1)$ voisins (également tiré d'une loi binomiale). Lorsqu'on permet à p de varier avec N, on peut obtenir des graphes creux en choisissant $p(N) = o(1)$, par exemple $p(N) = c/N$. Le degré des noeuds est alors tiré d'une loi de Poisson de moyenne c, et le nombre de liens est relié linéairement au nombre de noeuds. Grâce à la loi des grands nombres, on sait que la distribution de $M(p)$ est centrée sur sa moyenne $\langle M(p) \rangle$ lorsque $N \gg 1$, de sorte que les modèles microcanonique et canonique sont *quasiment* [6] équivalents dans la limite des grands N.

Modèle des configurations. [7] Le constat empirique ayant le plus contribué à la naissance de la science des réseaux est probablement l'observation que le degré des noeuds n'est pas distribué uniformément dans les réseaux réels [16, 168]. Le modèle ER ne saisit pas du tout cette propriété des réseaux réels ; en fait, les noeuds y sont statistiquement interchangeables, ayant un degré identique en moyenne [8]. Ainsi, il faut imposer de nouvelles propriétés si on veut obtenir un certain niveau de réalisme pour des réseaux générés aléatoirement.

[6] Voir [27, § VII] pour une discussion détaillée des conditions d'équivalence.

[7] *Note historique* : Le modèle de configuration (CM) a été proposé indépendamment dans plusieurs disciplines au tournant des années 1980. En écologie, on l'a utilisé comme hypothèse nulle [44], alors qu'il a été étudié en tant que généralisation des graphes de Erdős-Rényi en combinatoire [19, 20] et en théorie des graphes [26], voir [79, §1.3 et 1.4] pour une revue complète. En science des réseaux, on attribue le CM principalement à Mark E.J. Newman et ses collaborateurs [35, 178], car ils ont complètement caractérisé le modèle à l'aide des fonctions génératrices de probabilités (PGF) [240], bien adaptées aux multiples généralisations qui ont suivies [7, 9, 10]. Toutefois, il est important de souligner que Michael Molloy et Bruce A. Reed sont arrivés aux mêmes résultats quelques années plus tôt [156, 157], à l'aide d'une approche différente, plus près de celle de Erdős et Rényi [68]. Tous ces modèles peuvent être vus comme microcanoniques, car on y fixe la séquence des degrés exactement ; une version canonique du CM a été proposée et popularisée en mathématiques par Fan Chung et Linyuan Lu [40].

[8] Le modèle reste utile, par exemple comme *modèle nul*. Il ne s'agit toutefois pas d'un modèle générant des réseaux réalistes.

Le *modèle de configuration* (CM) permet de pallier ce manque de façon directe. À l'instar du modèle de ER, il existe de nombreuses versions du CM [79], qu'on regroupera ici en deux grandes catégories par souci de simplicité. Dans les versions microcanoniques du modèle, on associe chacun des éléments d'une séquence de degrés $\boldsymbol{k} = [k_1, k_2, ..., k_N]^\intercal$ à un noeud. Le modèle est alors défini comme la distribution uniforme sur l'ensemble $\Omega_{\mathrm{CM}}(\boldsymbol{k})$ des réseaux avec la séquence de degrés $\boldsymbol{k}$, i.e., $P(G) = 1/|\Omega_{\mathrm{CM}}(\boldsymbol{k})|$ si $\hat{\boldsymbol{k}}(G) = \boldsymbol{k}$, alors que $P(G) = 0$ autrement. On peut permettre ou interdire les boucles et les liens multiples, ce qui mène à différents ensembles $\Omega_{\mathrm{CM}}(\boldsymbol{k})$ et donc différentes versions du modèle [79].

Il est facile de vérifier que lorsque le plus grand élément de la séquence des degrés est d'ordre $o(N)$, les réseaux de l'ensemble sont creux, puisque le degré moyen doit nécessairement être sous-linéaire. En effet, on a alors :

$$\rho = \frac{\langle k \rangle}{N} + O(N^{-1}) \leq \frac{o(N)}{N} \to 0\,,$$

où on a d'abord remplacé $\langle k \rangle$ par sa borne supérieure $k_{\max} = O(N)$. Lorsque les réseaux générés par le modèle sont creux, on peut traiter les liens comme indépendants *de façon effective*, i.e., la valeur de a_{ij} ne nous informe pas sur la valeur de a_{rs} pour $i \neq r$ et $j \neq s$ [45]. Les détails mathématiques justifiant cette indépendance sont complexes, mais les conséquences sont simples. En effet, lorsque cette définition tient, on peut écrire la moyenne de a_{ij} dans l'ensemble comme

$$\langle a_{ij} \rangle = \frac{k_i k_j}{2M} + o(M)\,. \tag{1.7}$$

On arrive à cette formule en considérant le voisinage de v_i, qu'on visualise localement comme un noeud entouré de k_i demi-liens (i.e., des liens sans voisin). Pour accrocher v_i au réseau, on doit lui trouver k_i voisins, i.e., choisir k_i demi-liens parmi les $(2M - k_i)$ demi-liens restants. Puisque le noeud v_j possède k_j demi-liens, il sera sélectionné comme voisin avec probabilité $k_j/(2M - k_i)$. On utilise alors l'hypothèse d'indépendance des liens et une approximation pour obtenir l'Eq. (1.7). Cette version décorrélée est facile à traiter analytiquement, et est donc typiquement celle qui est utilisée en science des réseaux [170].

Ce raisonnement est poussé plus loin dans la version canonique du CM : l'équation (1.7) y est prise comme *définition* plutôt que comme conséquence. Ainsi, tout comme dans le cas du modèle ER, les liens sont indépendants (alors qu'ils sont corrélés dans la version microcanonique), et les degrés sont maintenant fixés *en moyenne* plutôt que directement. On souligne ces différences en utilisant une séquence des degrés *moyens* $\boldsymbol{\kappa} = [\kappa_1, \kappa_2, ..., \kappa_N]^\intercal$ plutôt qu'une séquence exacte $\boldsymbol{k}$. Cette définition concorde avec le modèle, puisqu'on vérifie facilement que si a_{ij} est tiré d'une distribution quelconque de moyenne $(\kappa_i \kappa_j)/(2\mu)$, alors

$$\langle k_i \rangle = \sum_j \langle a_{ij} \rangle = \sum_j \frac{\kappa_i \kappa_j}{2\mu} = \kappa_i \tag{1.8}$$

où les moyennes sont prises sur l'ensemble et où $\mu = \frac{1}{2}\sum_i \kappa_i$ est le nombre de liens espérés dans l'ensemble. Puisque les liens sont explicitement indépendants dans la variante canonique du CM, on peut écrire la probabilité $P(G)$ de G comme $P(G) = \prod_{i\leq j} \mathbb{P}(a_{ij})$, où le choix d'une distribution pour a_{ij} fixe $\mathbb{P}(a_{ij})$. Les réseaux réels étant pratiquement tous creux [168], on opte normalement pour la distribution de Poisson, pour ses propriétés analytiques [163, § II] [9]. On a alors

$$P(G) = \prod_{i<j} \frac{(\kappa_i\kappa_j/2\mu)^{a_{ij}} e^{-\frac{\kappa_i\kappa_j}{2\mu}}}{a_{ij}!} \,, \tag{1.9}$$

une expression qui se réduit approximativement à l'Eq. (1.5) lorsque $\kappa_i = \kappa \;\forall\, i$.

Modèle stochastique par blocs. [10] Une autre propriété quasi universelle des réseaux réels est leur tendance à contenir des sous-ensembles de noeuds *structurellement équivalents* [76, 170, 203], i.e., des ensembles de noeuds qui sont connectés de la même façon avec le reste du réseau [72, 236, 238] [11]. Le modèle ER reproduit bien cette propriété au niveau global, mais ne permet pas de variations : *tous* les noeuds y sont équivalents, sans exception (voir **Modèle de Erdős-Rényi** ci-haut). Un réseau généré par le CM contient aussi des ensembles de noeuds équivalents, déterminés par leur degré. Or, en général, il existe des classes d'équivalence au-delà de la séparation par degrés [116]. Ainsi, si on veut générer des réseaux aléatoires plus réalistes, on doit explicitement tenir compte de la séparation des noeuds en sous-ensembles.

Le *modèle stochastique par blocs* (SBM) atteint cet objectif en incorporant la notion d'équivalence statistique au sein même de sa définition. Tout comme pour les autres modèles introduits jusqu'à présent, il existe des versions canonique et microcanonique du SBM, ainsi que de nombreuses généralisations; on se concentre uniquement sur les variantes les plus simples du SBM (étudiées dans les chapitres 2–3) et on réfère le lecteur intéressé à [1] pour une revue récente [12]. Toutes les définitions du SBM débutent avec une partition des noeuds en q blocs, dénotée $\mathcal{B} = \{B_1, ..., B_q\}$, où les $\{B_r\}$ (les blocs) sont des sous-ensembles disjoints et exhaustifs de V, i.e.,

$$B_r \cap B_s = \emptyset \quad \forall\, r \neq s \qquad \text{et} \qquad \bigcup_{r=1}^{q} B_r = V \,.$$

[9] Ce choix permet également de faire coïncider le CM microcanonique au CM canonique, dans la limite $N \to \infty$ des réseaux creux [45]. Encore une fois, cette équivalence ne tient pas forcément pour toutes propriétés, voir [79].

[10] *Note historique* : Le modèle stochastique (SBM) est défini pour la première fois à la frontière de la sociologie et de la statistique (la *sociométrie*), par P.W. Holland, K.B Laskey et S. Leinhardt [105], voir également les références [72, 73] pour une mise en contexte historique et les références [106, 238], où toutes les idées nécessaires à la définition du SBM sont mises en place. Le SBM passe d'un modèle théorique à un modèle appliqué au tournant du millénaire, grâce aux avancées computationnelles de K. Snijders et T.A.B Nowicki [179, 224]. Le modèle est par la suite redécouvert sous d'autres noms et de façon indépendante dans plusieurs domaines. Par exemple, en informatique théorique, la bissection de graphes aléatoires est étudiée par plusieurs dans les années 1990 et 2000 [43]. On reconnaît éventuellement qu'il s'agit d'instances particulières du SBM [1]. En physique statistique, on montre qu'une sous-classe des verres de spins de dimension infinie [227] peut être aussi vue comme étant générée par le SBM [56, 109, 207], une connexion qui mène ultimement à la (riche) théorie moderne du SBM en tant qu'outil d'inférence [252].

[11] Équivalence stricte : $v_i \sim v_j$ si et seulement si $a_{ik} = a_{jk}\; \forall k$. Équivalence statistique : les caractéristiques moyennes du réseau sont maintenues sous permutation $\pi(B)$ des noeuds dans un sous-ensemble $B \subset V$.

[12] Notons en particulier l'existence d'une version corrigée pour les degrés, qui combine CM et SBM [13, 190].

Il est souvent pratique de référer à la partition à l'aide de l'application $\sigma : V \to \{1,...,q\}$ qui assigne l'indice d'un bloc à chaque noeud. Puisque la partition induit en quelque sorte des sous-réseaux en interactions, il convient de raffiner les mesures globales définies dans la section 1.1.1. Ainsi, en plus de caractériser G à l'aide du nombre total de noeuds N et de liens M, on dénote maintenant par n_r le nombre de noeuds dans le bloc B_r ($n_r = |B_r|$), et par m_{rs} le nombre de liens connectant les noeuds de la paire de blocs (B_r, B_s) :

$$m_{rs} = \sum_{i<j} a_{ij}\, \mathbb{I}[\sigma(i) = r \wedge \sigma(j) = s] \, , \tag{1.10}$$

où $\mathbb{I}(X)$ est la fonction indicatrice, égale à 1 lorsque l'énoncé X est vrai et à 0 sinon. Les bornes supérieures sur les m_{rs} sont, pour les graphes simples,

$$m_{rs}^{\max} = n_r n_s \quad \text{si } r \neq s \quad \text{et} \quad m_{rs}^{\max} = \binom{n_r}{2} \quad \text{si } r = s \, . \tag{1.11}$$

Par analogie au modèle ER, la version microcanonique du SBM [190, 192, 193, 195–197] est alors définie comme la distribution uniforme sur l'ensemble $\Omega(\boldsymbol{n}, \boldsymbol{m})$ des graphes de $\boldsymbol{n}$ noeuds (le vecteur des n_r), et étant connecté par précisément $\boldsymbol{m}$ liens (la matrice des m_{rs}). Puisqu'il existe $|\Omega(\boldsymbol{n}, \boldsymbol{m})| = \prod_{r \leq s} \binom{m_{rs}^{\max}}{m_{rs}}$ graphes, le SBM microcanonique place une probabilité $P(G) = 1/|\Omega(\boldsymbol{n}, \boldsymbol{m})|$ connue sur un graphe (partitionné) ayant la bonne matrice $\hat{\boldsymbol{m}}(G)$, et place une probabilité $P(G) = 0$ sur G si $\hat{\boldsymbol{m}}(G) \neq \boldsymbol{m}$ [190].

À l'instar des modèles introduits jusqu'à présent, la version canonique est obtenue en remplaçant la condition stricte sur la matrice $\boldsymbol{m}$ par des probabilités. On introduit la matrice $\boldsymbol{P}$ dont l'élément $p_{rs} \in [0,1]$ est la probabilité qu'un noeud $i \in B_r$ soit connecté à un noeud $j \in B_s$. Dans le modèle, chaque lien est considéré indépendant, de sorte que la probabilité de G est donnée par

$$P(G) = \prod_{i<j} (1 - p_{\sigma(i)\sigma(j)})^{1-a_{ij}} p_{\sigma(i)\sigma(j)}^{a_{ij}} \, . \tag{1.12}$$

Plus succinctement, on peut regrouper les termes par paires de blocs (B_r, B_s) comme à l'Eq. (1.6). On obtient alors

$$P(G) = \prod_{r \leq s} (1 - p_{rs})^{m_{rs}^{\max} - m_{rs}} p_{rs}^{m_{rs}} \, , \tag{1.13}$$

où m_{rs} est fonction de G et de la partition $\mathcal{B}$, tandis que les bornes supérieures $m_{rs}^{\max}$ sont déterminées uniquement par $\mathcal{B}$. Le nombre de liens entre les noeuds des blocs B_r et B_s est dans ce cas tiré d'une distribution binomiale de moyenne $m_{rs}^{\max} p_{rs}$. Ceci implique que le degré moyen d'un noeud dans B_r est approximativement donné par $\langle k_r \rangle = \sum_r p_{rs} n_s$ et que le degré moyen *global* est égal à $\langle k \rangle = \frac{2}{n} \sum_{r \leq s} m_{rs}^{\max} p_{rs}$. Ce degré global est également distribué selon la loi binomiale, de paramètre $\rho = \langle k \rangle / (n-1)$, de sorte que la distribution devient poissonienne si $\langle k \rangle = o(N)$ dans la limite $N \to \infty$.

Avant de passer au dernier modèle à l'équilibre[13], on note que le SBM contient le modèle d'Erdős-Rényi comme cas spécial. Cette correspondance peut être obtenue de deux façons.

[13] Sans notion *dynamique* dans leur définition.

D'une part en choisissant $q = 1$ (les définitions retombent trivialement sur celles du modèle ER), et d'autre part en considérant des matrices de connectivité $\boldsymbol{m}$ ou $\boldsymbol{P}$ quasi uniforme, données par $\boldsymbol{m} = C_1 \boldsymbol{J}_q + \boldsymbol{\epsilon}$ (microcanonique) et $\boldsymbol{P} = C_2 \boldsymbol{J}_q + \boldsymbol{\epsilon}$ (canonique), où $\boldsymbol{J}_q$ est une matrice $q \times q$ de 1, où C_1 et C_2 sont des constantes, et où les éléments de $\boldsymbol{\epsilon}$ sont petits devant le terme $\boldsymbol{J}_q$. En effet, pour des matrices de connectivités quasi uniformes, la signature des blocs est tellement faible que les modèles ER et SBM sont essentiellement contigus (i.e., statistiquement indiscernables). C'est la formalisation de cette observation qui sera la base de notre analyse du chapitre 3.

Entropie et ERGMs. [14] Ce n'est pas par hasard que tous les modèles présentés jusqu'à maintenant ont la même saveur mathématique : leurs définitions sont en fait des conséquences simples du principe de maximisation de l'entropie. Ainsi, en prenant le concept d'entropie maximale comme unique point de départ, on peut voir ces modèles comme des cas spéciaux d'une construction beaucoup plus générale, chose que nous allons maintenant montrer [15]. Rappelons d'abord que *l'entropie de Shannon* d'une distribution $P(G)$ sur un ensemble Ω est donnée par [48, 221]

$$S(P) = -\sum_{G\in\Omega} P(G)\log P(G) \tag{1.14}$$

où $P(G)$ est la probabilité de l'élément $G \in \Omega$ [16]. L'entropie caractérise l'incertitude associée à la distribution P. Lorsque $S(P) \approx 0$, on peut prédire avec quasi certitude l'élément qui sera tiré de Ω ; à l'inverse, $S(P) \gg 0$ indique un grand niveau de surprise.

Nous allons maintenant chercher la famille de modèles qui maximise cette entropie sous contraintes. On peut voir ce principe comme un rasoir d'Ockham statistique. Il s'agit d'une opérationnalisation de l'idée qu'un modèle aléatoire, i.e., une distribution $P(G)$ sur un ensemble de graphes Ω, doit être le plus imprévisible possible, tout en respectant les contraintes observées / imposées. Formellement, on cherche à maximiser $S(P)$ tout en fixant (i) une normalisation $\sum_{G\in\Omega} P(G) = 1$, et (ii) une série de k contraintes $\boldsymbol{\theta}(G) = \{\theta_r(G)\}_{1,..,k}$. Ces contraintes peuvent être fixées exactement, i.e.,

$$\theta_r(G) = \Theta_r \quad \forall\, G \in \Omega\,, \tag{1.15}$$

ou en moyenne, i.e.,

$$\sum_{G\in\Omega} P(G)\theta_r(G) = \Theta_r \tag{1.16}$$

[14] *Note historique* : Tout comme le SBM, les modèles exponentiels de graphes aléatoires (ERGM) sont introduits au tournant des années 1980 en sociométrie [106]. Le cadre conceptuel est complété quelques années plus tard par Frank et Strauss, qui placent les ERGMs dans le contexte plus large des modèles log-linéaires pour graphes [80]. Les modèles log-linéaires (généraux) sont également connus sous les noms de *maximum-entropy models, Gibbs Models,* ou encore de *Markov Random Fields* [153] ; en physique statistique, on doit principalement ce type de description à Jaynes [112]. La description originale des ERGMs étant extrêmement générale, la plupart des avancées en inférence et modélisation de graphes aléatoires ont des ramifications pour ces modèles ; voir par exemple la référence [45] qui traite de graphes aléatoires (quasiment) uniquement en termes d'ERGM .

[15] Cette sous-section est inspirée des chapitres 3 et 4 de la référence [45].

[16] La base du logarithme importe peu. On optera pour e, de sorte que $S(P)$ est calculé en *nats*.

pour un choix de paramètres $\boldsymbol{\Theta} = [\Theta_1, \ldots, \Theta_k]$, qu'on suppose dans $\mathbb{R}$ par souci de simplicité (tout comme $\boldsymbol{\theta}$). Dans les deux cas, la solution est donnée par le calcul des variations [45, 112] ; on doit trouver la distribution P qui maximise l'entropie tout en respectant l'Eq. (1.15) ou l'Eq. (1.16).

Commençant par le cas des contraintes exactes, on a trivialement que si $\theta_r(G) \neq \Theta_r$ pour au moins un $r = 1, .., k$, alors $P(G) = 0$. On peut donc *définir* l'ensemble Ω comme l'ensemble des graphes respectant $\boldsymbol{\theta} = \boldsymbol{\Theta}$. Afin de trouver la distribution sur les graphes dans Ω, on cherche alors le point fixe de

$$\frac{\delta}{\delta P}\left[S(P) + \lambda\left(\sum_{G\in\Omega} P(G) - 1\right)\right] = 0\,, \tag{1.17}$$

où λ est un multiplicateur de Lagrange. On trouve que la distribution P est donnée par $\log P = \lambda - 1$. Utilisant la seconde équation d'optimisation (la normalisation), on obtient $\lambda = 1 - \log|\Omega|$, et donc

$$P(G) = \begin{cases} 1/|\Omega| & \text{si } \boldsymbol{\theta}(G) = \boldsymbol{\Theta}, \\ 0 & \text{sinon,} \end{cases} \tag{1.18}$$

où $|\Omega|$ est le nombre de graphes respectant les contraintes. On aura reconnu directement que cet ensemble est essentiellement un ensemble microcanonique de réseaux.

La démarche est similaire dans le cas avec contraintes moyennes. La différence principale réside dans le fait qu'on ne peut plus invoquer $P(G) = 0$ et redéfinir Ω ; la distribution donne *a priori* un poids non nul (quoique potentiellement très petit) à tous les graphes dans Ω. On impose donc plutôt les contraintes moyennes à l'aide des k multiplicateurs de Lagrange $\boldsymbol{\beta} = \{\beta_r\}_{r=1,..,k}$, interprétés comme des températures inverses :

$$\frac{\delta}{\delta P}\left[S(P) + \lambda\left(\sum_{G\in\Omega} P(G) - 1\right) + \sum_{r=1}^{k}\beta_r\left(\sum_{G\in\Omega} P(G)\theta_r(G) - \Theta_r\right)\right] = 0\,. \tag{1.19}$$

La distribution P d'intérêt est alors donnée par

$$\log P(G) = \lambda - 1 + \sum_{r=1}^{k}\beta_r\theta_r(G) = 0\,.$$

Utilisant le fait que P doit être normalisée (première contrainte), on obtient la forme générale

$$P(G) = \frac{e^{\sum_{r=1}^{k}\beta_r\theta_r(G)}}{Z(\boldsymbol{\beta})}\,, \qquad Z(\boldsymbol{\beta}) = \sum_{G\in\Omega}\exp[\sum_r \beta_r\theta_r(G)]\,, \tag{1.20a}$$

où les k autres multiplicateurs sont déterminés par les k équations de contraintes

$$\sum_{G\in\Omega}\frac{e^{\sum_{r=1}^{k}\beta_r\theta_\ell(G)}\theta_r(G)}{Z(\boldsymbol{\beta})} = \Theta_\ell\,. \tag{1.20b}$$

Les équations (1.20) définissent un ensemble analogue à l'ensemble canonique de la physique statistique, connu principalement sous le nom de modèle exponentiel de graphes aléatoires (ERGM).

Les modèles des sections précédentes ont manifestement la même saveur que les ERGMs, car on y fixe *uniquement* la densité (ER), les degrés (CM) et la structure en blocs (SBM), laissant le reste au hasard. On devrait donc s'attendre à pouvoir réécrire leurs probabilités $P(G)$ sous la forme (1.18) dans le cas microcanonique, et sous la forme (1.20) dans le cas canonique. En pratique, cette correspondance est triviale à établir dans le cas microcanonique, car les définitions concordent parfaitement—aucun calcul n'est requis. Le cas canonique est plus ardu. En effet, pour obtenir $P(G)$ pour un choix de contraintes $\boldsymbol{\theta}$, il faut calculer les températures inverses $\boldsymbol{\beta}$. La méthode standard pour y arriver consiste à obtenir l'entropie libre $\log Z(\boldsymbol{\beta})$ en forme fermée, ce qui nous permet alors de calculer

$$\frac{\partial}{\partial \beta_\ell} \log Z(\boldsymbol{\beta}) \equiv \sum_{G\in\Omega} \frac{e^{\sum_{r=1}^{k} \beta_r \theta_r(G)} \theta_\ell(G)}{Z(\boldsymbol{\beta})} = \Theta_\ell \,, \tag{1.21}$$

i.e., une autre expression pour les équations de contraintes.

Afin d'illustrer le genre de calcul complexe que cette approche implique, considérons les cas simples du modèle de Erdős-Rényi et du SBM [17]. La contrainte d'intérêt pour le modèle de ER est le nombre moyen de liens $\langle M\rangle$, calculée comme $M = \sum_{i<j} a_{ij}(G)$ pour un réseau G donné de l'ensemble. Notre point de départ est donc la probabilité (1.20a), ici donnée par

$$P(G,\beta) = \frac{\exp\left(\beta \sum_{i<j} a_{ij}(G)\right)}{Z(\beta)} \qquad \text{et} \qquad Z(\beta) = \sum_{G\in\Omega} \exp\left(\beta \sum_{i<j} a_{ij}(G)\right) \,, \tag{1.22}$$

où la température inverse β est pour l'instant inconnue, et doit être choisie de façon à ce que $\sum_{G\in\Omega} P(G,\beta) M(G) = \langle M\rangle$. Pour trouver β, on utilise d'abord l'identité $\sum_{G\in\Omega} \prod_{i<j} f(a_{ij}) = \prod_{i<j} \sum_{a_{ij}\in\{0,1\}} f(a_{ij})$, valide pour les graphes simples [18], et on obtient

$$\log Z(\beta) = \log \prod_{i<j} \sum_{a_{ij}\in\{0,1\}} e^{\beta a_{ij}} = \binom{N}{2} \log(1+e^{\beta}). \tag{1.23}$$

L'équation (1.21) donne alors

$$\binom{N}{2} \frac{e^{\beta}}{1+e^{\beta}} = \langle M\rangle \qquad \Longrightarrow \qquad \beta = \log \frac{\rho}{1-\rho} \tag{1.24}$$

où $\rho := \langle M\rangle / \binom{N}{2}$ est la densité moyenne dans l'ensemble. Avec la température inverse et la fonction de partition en main, on peut finalement retourner à l'Eq. (1.22), qu'on réécrit comme

$$P(G,\beta) = \prod_{i<j} \rho^{a_{ij}} (1-\rho)^{1-a_{ij}} \tag{1.25}$$

[17] On omet la démonstration que le CM est un ERGM, obtenue en imposant N contraintes sur les degrés moyens $\{\kappa_i\}_{i=1,..,N}$. Le lecteur intéressé peut se référer à [45, p.37] pour une démarche complète.

[18] La somme sur Ω peut alors être vue comme une somme sur les matrices binaires $\boldsymbol{A} \in \{0,1\}^{N\times N}$.

en utilisant $e^{\beta \sum_{i<j} a_{ij}} = \prod_{i<j} (e^{\beta})^{a_{ij}} (1)^{1-a_{ij}}$. On reconnaît que cette dernière équation est identique à l'Eq. (1.5) ci-haut, ce qui démontre que le modèle d'Erdős-Rényi fait partie de la famille des ERGMs, par équivalence mathématique.

Un calcul similaire permet d'en faire autant pour le SBM, car ce dernier peut être vu comme une superposition de graphes ER. On passe simplement d'une seule contrainte sur $\sum_{i<j} a_{ij}$ à $\binom{q}{2} + q$ contraintes sur le nombre de liens $\langle m_{rs} \rangle$ entre les noeuds des blocs (r,s) ; le reste du raisonnement n'est pas affecté. On peut ainsi reprendre l'entropie libre de l'Eq. (1.23) et écrire

$$\log Z(\boldsymbol{\beta}) = \sum_{r,s} m_{rs}^{\max} \log(1 + e^{\beta_{rs}}) \,. \tag{1.26}$$

En guise de vérification, on peut montrer qu'on retombe sur le cas ER lorsque $q = 1$ (une seule contrainte), où $m_{rs}^{\max} = \binom{N}{2}$ (voir section précédente). Solutionnant pour $\boldsymbol{\beta}$ à l'aide de (1.21), on obtient

$$\beta_{rs} = \log \frac{2\rho_{rs}}{1 - 2\rho_{rs}} \tag{1.27}$$

où $\rho_{rs} = \langle m_{rs} \rangle / m_{rs}^{\max}$, ce qui mène directement à une équation de la même forme que (1.13). Le SBM canonique est donc également un cas spécial des ERGMs.

Attachement préférentiel.[19] L'indépendance des liens est au coeur des modèles introduits jusqu'à présent. Elle est soit imposée explicitement (modèles canoniques) ou obtenue dans la limite des réseaux creux (modèles microcanoniques). Or, il s'agit d'une hypothèse forte, assurément erronée dans plusieurs cas réels[20]. Avant de conclure notre survol des ensembles aléatoires de réseaux, on se doit d'aborder au moins un modèle où cette hypothèse n'est pas utilisée. On introduit donc un modèle classique de la science des réseaux, celui de l'attachement préférentiel (PA) [16]. Contrairement aux modèles précédents, le modèle PA ne donne pas directement une distribution $P(G)$ sur les réseaux. Il s'agit plutôt d'un modèle dit mécaniste, défini directement en termes d'un processus de construction aléatoire de G (en d'autres mots, la distribution $P(G)$ est une conséquence du processus et non l'inverse). Cette construction permet d'introduire des corrélations sans avoir à les traiter explicitement. La version de Krapivsky et Redner du modèle PA procède comme suit [128]. On considère

[19] *Note historique* : Le modèle d'attachement préférentiel (PA) est introduit en science des réseaux par R. Albert et A.-L. Barabási en 1999 [16]. Les conséquences de ce modèle sont rapidement déterminées dans les années qui suivent, entre autres par P.L Krapivsky et S. Redner et leurs collaborateurs [126, 128, 129], ainsi que S. Dorogovtsev, J.F.F Mendes [63], voir la revue [61]. Le modèle PA construit sur l'idée du *rich-get-richer*, déjà très répandue en complexité. On pense notamment aux travaux de Pólyà, de Yule et de Gibrat, qui précèdent la naissance de la science de la complexité (1920–1940), et à ceux de Simons, un des pères fondateurs du domaine [222].

[20] Il ne s'agit toutefois pas d'un constat d'échec pour ces méthodes. En statistique, les simplifications amenées par l'indépendance permettent souvent d'arriver à des conclusions utiles, même quand cette hypothèse n'est pas justifiée [36]. De même, il existe des exemples connus de modèles non corrélés utiles dans des cas avec corrélations. Voir, par exemple : la référence [180] qui montre que le SBM offre une approximation de la structure de *tous* les réseaux ; la référence [152] qui montre que le CM et ses généralisations mènent à de bonnes prédictions de dynamique sur réseaux réels ; ou encore le chapitre 2 où on montre que plusieurs méthodes de détection communautaire peuvent être réécrites comme un problème d'inférence SBM.

d'abord un graphe initial G_0 comme le point de départ d'un processus en temps discret $t = 1...,T$ associé à une suite de graphes

$$G_0 \to G_1 \to \ldots \to G_T \ .$$

À chaque pas de temps discret, on sélectionne un noeud v_i de G_{t-1} au hasard et on attache un *nouveau* noeud v_j à v_i, à l'aide d'un lien dirigé $v_j \to v_i$. Afin d'introduire des inhomogénéités dans la distribution des degrés du réseau, cette sélection est faite avec un poids

$$p_i(t) = \frac{k_i^\gamma(t-1)}{\sum\limits_{i \in V(G_{t-1})} k_i^\gamma(t-1)} \tag{1.28}$$

pour le noeud v_i, où $k_i(t)$ est son degré au temps t et où $\gamma \in \mathbb{R}$ est l'exposant du *noyau d'attachement* non-linéaire $g(k) = k^\gamma$. Lorsque $\gamma = 0$, les nouveaux noeuds sont attachés uniformément aux noeuds déjà existants, tandis que $\gamma > 0$ favorise les inhomogénéités. Ainsi $\gamma \in (0,1)$ mène à des réseaux dont la distribution des degrés est une loi de puissance couplée à une décroissance exponentielle, i.e., $p_k \propto k^{-a} \exp[-f(k)^b]$. Cette distribution devient une loi de puissance parfaite lorsque $\gamma = 1$, i.e., lorsque le modèle correspond au modèle PA classique [15]. Pour $\gamma > 1$ une fraction extensive des liens s'agrège dans un *condensat* [126] ; ce phénomène est exacerbé à $\gamma = 2$, où quasiment tous les noeuds sont attachés à un seul noeud central.

Si la probabilité d'attachement (1.28) est facile à écrire, ce n'est pas le cas de $P(G)$, parce qu'il faudrait sommer sur l'ensemble des chemins qui mènent à un même graphe [45, § 9]. Il n'est donc pas facile de démontrer l'existence de corrélations en calculant directement $P(A_{ij}, A_{jk}) \neq P(A_{ij})P(A_{jk})$, où A_{ij} est la variable aléatoire associée au lien (v_i, v_j). Cependant, une rapide inspection du processus lui-même nous confirme que le modèle génère des réseaux corrélés. Par exemple, on sait que l'âge d'un noeud est relié à son degré [16], ce qui introduit une corrélation degré-degré importante, puisque les plus vieux noeuds partagent plus de connexions, par construction [45]. Ainsi la probabilité $P(k,k')$ de trouver des noeuds de degré k et k' aux deux extrémités d'un lien choisi uniformément n'est pas proportionnelle à $P(k)P(k')$, où $P(k)$ est la probabilité qu'un noeud choisi uniformément soit de degré k. Ce genre de corrélations (il y en plusieurs autres [61]) nous assurent que PA est différent, par exemple, du CM avec la même distribution des degrés. Le modèle PA peut donc être utilisé pour reproduire les corrélations de systèmes réels qui sont ignorées par les ERGMs, par exemple.

On note que les hypothèses menant au modèle PA (présenté ci-haut) ne sont pas toujours pertinentes et applicables ; il ne reproduit donc pas forcément tous les phénomènes jouant un rôle important dans la formation des réseaux réels. Mais, en raison de sa simplicité, ce modèle peut agir comme substrat pour une foule de généralisations. On peut par exemple : ignorer la direction des liens ; attacher m noeuds à chaque pas de temps [16] ; considérer un attachement au niveau de superstructures [97, 102, 248] ; ajouter des préférences cachées pour les noeuds [21]. On considérera un exemple de généralisation au chapitre 4, où on introduit une version non dirigée et densifiée du modèle, dans laquelle un événement d'attachement peut impliquer deux noeuds existants, avec probabilité $1-p \in [0,1)$.

$\therefore$

Ce dernier modèle complète notre survol des ensembles aléatoires de réseaux complexes : modèle de Erdős-Rényi (ER), modèle des configurations (CM), modèle stochastique par blocs (SBM), modèles exponentiels de graphe aléatoires (ERGMs), et attachement préférentiel (PA). On a vu que les trois premiers modèles—tous formulés à l'équilibre—peuvent être considérés comme maximalement entropiques, unifiés sous l'égide des ERGMs et de la distribution uniforme. On a conclu la section avec un modèle mécaniste (PA), fondamentalement différent car formulé comme un processus stochastique hors d'équilibre [21]. On a vu que ce modèle génère des réseaux corrélés, qui contiennent des inhomogénéités au niveau de la distribution des degrés. Si on a introduit ces modèles, c'est parce qu'ils seront le point de départ de notre exploration de l'inférence statistique sur réseaux, qu'on aborde maintenant plus en détails.

1.2 Inférence statistique et réseaux

Au sens large, on désigne par *inférence statistique* l'ensemble des méthodes numériques et analytiques permettant de *déduire* les propriétés d'un processus *stochastique*, à partir d'observations générées par le processus en question. Puisqu'une part de hasard intervient dans la génération d'observations, ces déductions sont nécessairement incertaines et exprimées dans le langage des probabilités et des modèles aléatoires. Concrètement, l'inférence statistique s'attaque à des problèmes comme l'estimation de paramètres, le test d'hypothèses, ou encore l'estimation d'intervalles [36]. Ces problèmes sont habituellement introduits en terme de distributions simples sur $\mathbb{R}$ ou $\mathbb{N}$ (e.g., estimation de la moyenne d'une loi gaussienne), mais ont des extensions naturelles à des espaces discrets arbitraires. Parmi ceux-ci, on compte évidemment des espaces de réseaux, de sorte qu'on peut parler d'inférence statistique pour réseaux.

[21] En tant que processus dynamique aléatoire.

1.2.1 Modèles nuls

I couldn't claim that I was smarter than sixty-five other guys
but the average of sixty-five other guys, certainly !
—Richard Feynman

On a déjà abordé les modèles nuls brièvement au début de la section 1.1.2, pour motiver l'introduction des ensembles statistiques de réseaux. On y revient formellement, maintenant armé d'un large éventail de modèles aléatoires nous permettant d'illustrer ses principes plus directement. Le point de départ d'une analyse par modèle nul d'un réseau G^* est (a) une propriété x et (b) un modèle aléatoire $\mathcal{M}$ qu'on contrôle à l'aide de paramètres $\boldsymbol{\theta}$. L'algorithme est généralement le suivant [29, 79, 182, 250] :

1. on calcule la propriété $x(G^*)$;
2. on génère des réseaux aléatoires $G \sim \mathcal{M}$ où les paramètres $\boldsymbol{\theta}$ du modèle $\mathcal{M}\big(\boldsymbol{\theta}(G)\big)$ sont fixés par la structure de G^* [22] ;
3. on calcule la distribution de $X(G)$—la variable aléatoire associée à x—, numériquement ou analytiquement ;
4. on compare $x(G^*)$ à la distribution de $X(G)$.

Si $x(G^*)$ est loin du centre de la distribution de X, alors on conclut que la propriété x n'est pas *expliquée* par les propriétés utilisées pour construire $\mathcal{M}$—il s'agit d'un modèle inadéquat. Autrement, si $x(G^*)$ est près de la valeur attendue, alors ont dit que le modèle $\mathcal{M}$ *explique* G^*. On peut voir cette méthode d'analyse comme une application du concept de «test de position» (*location test*). Ainsi, lorsque X est distribuée normalement [23], on peut quantifier la position de l'observable $x(G^*)$ à l'aide du nombre d'écarts-types à la moyenne.

Modèles nuls : exemple

Pour illustrer le concept de modèles nuls, on considère un exemple concret, soit un réseau des interactions sociales de dauphins vivant dans un fjord de Nouvelle-Zélande, obtenu à l'aide d'observations menées sur une période de 6 ans [145]. Le réseau, qui comporte 62 dauphins (noeuds) et 159 relations (liens), est montré dans la Fig. 1.1. Puisqu'il s'agit d'un réseau social animal, des questions scientifiques d'intérêt sont, par exemple, «*est-ce que les dauphins populaires préfèrent interagir entre eux ?*», ou encore «*est-ce que le cercle social des dauphins est*

[22] Les modèles de type microcanonique sont souvent préférés aux modèles canoniques pour ce genre de comparaison, car ils sont dits non paramétriques. Un modèle microcanonique est complètement spécifié par une série de propriétés $\boldsymbol{\Theta}(G^*)$ du réseau observé, tandis qu'on doit trouver la valeur de paramètres cachés dans le cas canonique (ce qui peut être difficile lorsque les contraintes sont complexes). Toutefois, le choix n'est pas automatique, car il est souvent plus aisé d'échantillonner un modèle canonique (voir, e.g., la méthode très flexible de la référence [74]). De plus, on peut vouloir permettre une certaine variation des contraintes, par exemple lorsque le réseau est bruité et que ses statistiques sont incertaines.

[23] C'est souvent le cas pour des propriétés réseaux globales, à cause du théorème central de la limite.

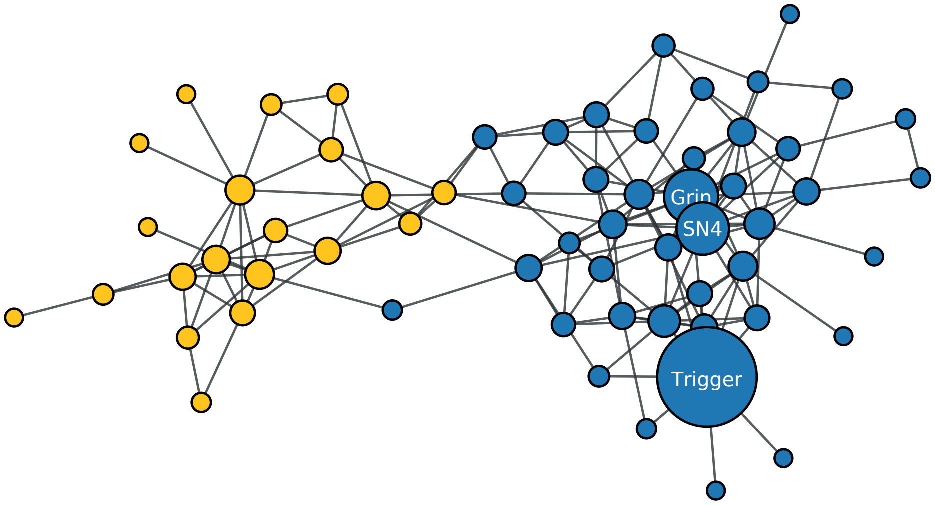


FIGURE 1.1 – Réseau des interactions sociales d'un groupe de dauphins [145]. La taille des noeuds est proportionnelle à leur degré (à l'exception des noeuds annotés), tandis que les couleurs représentent l'appartenance aux groupes identifiés par le SBM microcanonique classique [190, 194]. Les groupes sont robustes dans le sens où la même division est identifiée par plusieurs algorithmes différents, voir par exemple les références [176, 212].

généralement fermé ?». Les mesures réseaux permettant de répondre à ces questions sont, respectivement, le coefficient d'assortativité $r \in [-1,1]$, qui quantifie la propension globale des noeuds à se connecter à des noeuds de degré similaire, et le coefficient d'agrégation moyen $C \in [0,1]$, donné par la fraction des triangles qui sont fermés (i.e., la mathématisation du dicton «*les amis de mes amis sont mes amis*»). Leurs définitions sont [45, 166]

$$r = \frac{\sum_{k,k'} kk' w_{kk'} - \left(\sum_k k w_k\right)^2}{\sum_k k^2 w_k - \left(\sum_k k w_k\right)^2} \,, \tag{1.29}$$

où $w_{kk'}$ est la fraction des liens connectant des noeuds de degré (k,k') et $w_k = \sum_{k'} w_{kk'}$, et

$$C = \sum_{i=1}^{n} \frac{2t_i}{k_i(k_i-1)} \,, \tag{1.30}$$

où t_i est le nombre de paires de voisins connectées de v_i [i.e., nombre de paires (v_i, v_j), (v_i, v_k) tel que $(v_j, v_k) \in E(G)$].

On trouve que le coefficient d'agrégation du réseau de dauphins est grand ($C \approx 0.26 \gg 0$), tandis que le coefficient d'assortativité est quasi nulle ($r \approx -0.04$). D'un point de vue purement réseau, ces quantités nous indiquent que les voisins d'un noeud sont souvent connectés entre eux, et que le degré d'un noeud n'est pas un bon indicateur du degré de ses voisins. D'un point de vue sociologique, elles nous indiquent qu'un phénomène de fermeture triadique (*triadic closure*) joue probablement un rôle dans ce réseau (la présence des liens $a - b$ et

$a-c$ suggère la présence d'un lien $b-c$), et qu'il y a peu ou pas de préférence pour le statut social, tel que quantifié par le degré.

Il reste à vérifier si ces observations peuvent être expliquées par des propriétés simples du réseau. Pour ce faire, on compare le réseau observé à des réseaux aléatoires équivalents, générés à l'aide des trois modèles à l'équilibre introduit à la section 1.1.2 : le modèle ER, le CM et le SBM. On se limite aux versions canoniques, pour lesquelles le choix de paramètres est plus simple. Les paramètres des deux premiers modèles (ER, SBM) sont déterminés par observation directe du réseau. Dans le cas du modèle ER, on demande que les réseaux $\{G\}$ de l'ensemble aient précisément $N=62$ noeuds et $M=159$ liens, tandis que dans le cas du CM, on demande que leurs séquences de degrés soient toutes identiques à celle du réseau de dauphins, montrée à la Fig. 1.2 (c). Le choix de paramètres est plus difficile pour le SBM, car l'assignation à des blocs n'est pas une propriété observable directement. On trouve donc d'abord les blocs $\hat{\mathcal{B}}$ et les valeurs $\hat{m}_{rs}$ qui décrivent le mieux le réseau, en maximisant[24]

$$\log P(G) = -\sum_{r\leq s} \log \binom{m_{rs}^{\max}}{m_{rs}},$$

le logarithme de la vraisemblance du modèle. On génère ensuite des réseaux aléatoires à partir du SBM microcanonique de paramètres $\hat{\boldsymbol{m}}$ et $\hat{\mathcal{B}}$ inférés.

Les résultats de l'analyse par modèles nuls apparaissent aux figures 1.2 (a, b). On constate d'abord que tous les modèles aléatoires produisent des distributions de coefficient d'assortativité centrée sur une petite valeur de r, légèrement négative. Même si les modèles plus complexes comme le CM et le SBM mènent à des distributions plus concentrées sur la valeur observée r^*, on constate que même le modèle le plus simple (ER) suffit à expliquer l'absence de corrélations degré-degré. Ainsi, on peut dire que la valeur observée est non significative, dans le sens ou la plupart des réseaux de même densité et taille que G^* ont une assortativité faible, voir nulle. Cette observation est expliquée par le bruit.

On arrive à une conclusion différente lorsqu'on analyse le coefficient d'agrégation [voir Fig. 1.2 (b)]. En effet, on note qu'aucun modèle aléatoire ne réussit à générer des réseaux avec un niveau d'agrégation adéquat (i.e., comparable au niveau observé). Les modèles sans structure mésoscopique mènent à des coefficients particulièrement bas (en fait asymptotiquement nul pour des réseaux creux [170]). Le SBM fait mieux car, dans le cas à l'étude, la structure en blocs force la plupart des liens à se concentrer dans des sous-graphes plus petits, nécessairement plus denses et agrégés. [c.f. Fig. 1.2 (d)]. Mais puisqu'on mesure un coefficient d'agrégation encore plus grand pour le réseau réel, on doit conclure que même le SBM est un mauvais modèle : un principe encore plus fort que la séparation en communautés est nécessaire pour expliquer le niveau d'agrégation observé. À la lumière de ces observations,

[24] On utilise également un critère d'information pour sélectionner le nombre de blocs q. Voir la Sec. 1.3 et les chapitres 2–3 pour plus de détails.

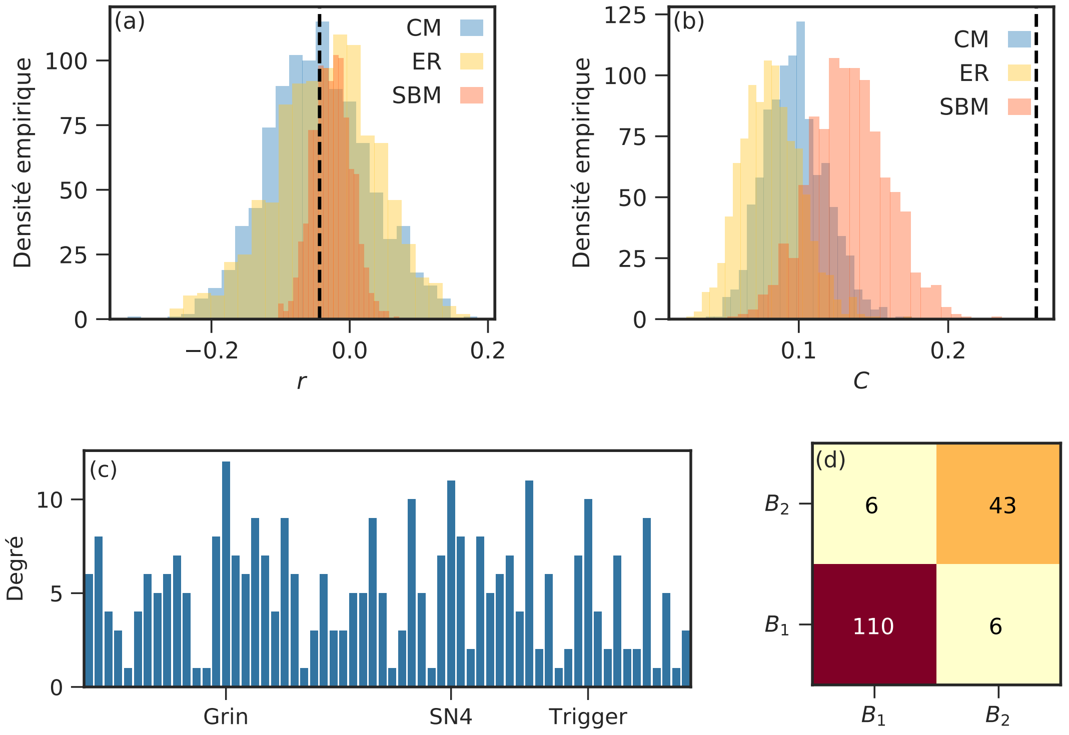


FIGURE 1.2 – Distribution des coefficients (a) d'assortativité et (b) d'agrégation, dans des modèles associés au réseau de dauphins de la référence [145]. On utilise la version microcanonique des modèles nuls, décrite dans la section 1.1.2. La valeur réelle des quantités est montrée à l'aide d'une ligne verticale noire ($r = -0.044$, $C = 0.26$). Les distributions sont calculées à partir sur 1 000 échantillons de chaque modèle aléatoire [79, 90, 194]. Les paramètres des modèles sont : (ER) le nombre de noeuds et de liens ; (CM) la séquence de degrés montrée en (c) ; et (SBM) un vecteur de taille de blocs $\boldsymbol{n} = [42, 20]$ et une matrice de liens $\boldsymbol{m}$, montrée en (d).

on ne peut pas rejeter l'hypothèse qu'il y a une tendance à la fermeture triadique dans le réseau social des dauphins. Un modèle où ce genre de corrélation apparaît explicitement pourrait nous permettre de tester cette hypothèse encore plus sérieusement.

1.2.2 Estimation de paramètres

Dans la section précédente, on a vu que les modèles nuls permettent de vérifier si une propriété $y(G^*)$ d'un réseau réel G^* explique une autre propriété $x(G^*)$. Pour ce faire, on a généré des réseaux artificiels $\{G\}$ qui imitent G^* (en fixant $y(G)$), et on a analysé la distribution de X pour ces réseaux. Ce faisant, on a utilisé la direction dite «vers l'avant» (*forward direction*) des modèles aléatoires, i.e., on a construit des réseaux artificiels à partir de modèles $\{\mathcal{M}\}$. On se penche maintenant sur le problème inverse (*backward direction*). Dans le problème inverse, on suppose[25] qu'un mécanisme a généré le réseau observé G^* ; la question d'intérêt devient alors : «*Quels sont les paramètres du mécanisme, sachant qu'on a observé G^* ?*»

Puisqu'on suppose que la structure est générée aléatoirement, et puisque les paramètres sont fonction de la structure, cette approche revient à voir les paramètres comme des variables aléatoires qu'on doit caractériser. Ce point de vue fructueux permet, par exemple, de postuler un mécanisme ayant mené à la création du réseau, et d'utiliser l'inférence statistique pour extraire de l'information de la structure observée, maintenant vue comme une signature dudit mécanisme (c.f. les chapitres 2 à 4). Il n'est donc pas surprenant que le problème d'estimation soit au coeur de l'inférence statistique réseau. Si le travail d'estimation de paramètres peut paraître trivial lorsqu'on considère des modèles microcanoniques simples tel le modèle ER ou le CM (il suffit d'inspecter G^* pour les trouver), le problème général est en fait très complexe. C'est pourquoi il s'agit d'un champ d'études extrêmement actif [110, 252].

Dans le cadre de cette thèse, on adopte un point de vue bayésien de l'inférence[26]. Notre point de départ sera donc la formule de Bayes

$$P(\boldsymbol{\theta}|G^*) = \frac{P(G^*|\boldsymbol{\theta})P(\boldsymbol{\theta})}{P(G^*)} \tag{1.31}$$

où $P(G^*|\boldsymbol{\theta})$ est la vraisemblance du réseau observé, $P(\boldsymbol{\theta})$ est une probabilité a priori sur les paramètres, $P(G^*)$ est l'évidence associée au modèle, et où $P(\boldsymbol{\theta}|G^*)$ est la distribution a posteriori sur les paramètres. On écrit la formule de Bayes pour un seul réseau G^* car on dispose généralement d'un échantillon de taille $n = 1$; il n'y a qu'un seul Internet, un seul *Facebook*, un seul réseau de distribution de courant, etc. Toutefois, si plusieurs réseaux sont disponibles, il est facile de corriger l'Eq. (1.31)—et les développements qui suivent—pour prendre toutes les observations en compte.

La première étape de toute démarche d'inférence bayésienne consiste à calculer la distribution a posteriori sur les paramètres $P(\boldsymbol{\theta}|G^*)$, pour une paire réseau–modèle $(G^*, \mathcal{M})$. Trois éléments entrent dans le calcul de cette probabilité :

1. la vraisemblance $P(G^*|\boldsymbol{\theta})$, qui quantifie l'information apportée par la structure ;
2. la probabilité a priori $P(\boldsymbol{\theta})$, qui permet d'incorporer une connaissance préalable de G^* ;
3. l'évidence associée au modèle $P(G^*)$.

Le troisième élément est toutefois redondant, car l'évidence $P(G^*)$ est complètement déterminée par la probabilité a priori et la vraisemblance. En effet, utilisant la définition des

[25] Il n'est pas nécessaire de faire ce choix aveuglément. On peut par exemple utiliser le test du ratio de vraisemblance, couplé à un critère d'information, pour choisir un bon modèle parmi une série de modèles.

[26] On peut développer une théorie de l'estimation équivalente dans le cadre plus classique de la statistique «fréquensiste», voir par exemple les chapitres 7 et 9 de la référence [36]. On choisit le point de vue bayésien pour suivre les conventions du domaine, mais également parce que la théorie bayésienne de l'inférence est plus proche des mathématiques de la physique statistique que la théorie classique équivalente [112, 153, 252].

probabilités conditionnelles, on observe que

$$P(G^*) = \begin{cases} \sum_{\{\boldsymbol{\theta}\}} P(G^*,\boldsymbol{\theta}) = \sum_{\{\boldsymbol{\theta}\}} P(G^*|\boldsymbol{\theta})P(\boldsymbol{\theta}) & \text{(paramètres discrets) ,} \\ \int P(G^*,\boldsymbol{\theta})d\boldsymbol{\theta} = \int P(G^*|\boldsymbol{\theta})\rho(\boldsymbol{\theta})d\boldsymbol{\theta} & \text{(paramètres continus).} \end{cases} \tag{1.32}$$

On peut également déterminer $P(G^*)$ par normalisation de la distribution a posteriori. Il suffit donc de choisir un modèle pour G^*, calculer sa vraisemblance, et de choisir une probabilité a priori pour pouvoir calculer la distribution a posteriori $P(\boldsymbol{\theta}|G^*)$. Une fois qu'on a obtenu la distribution $P(\boldsymbol{\theta}|G^*)$, on peut passer à l'estimation des paramètres $\hat{\boldsymbol{\theta}}$ du modèle. Plusieurs stratégies sont possibles [252].

La stratégie la plus directe consiste à maximiser la distribution a posteriori (MAP), i.e., à prendre

$$\hat{\boldsymbol{\theta}}^{\mathrm{MAP}} = \mathrm{argmax}_{\boldsymbol{\theta}}\, P(\boldsymbol{\theta}|G^*)\,. \tag{1.33}$$

On choisit ainsi les paramètres les plus probables, à partir de notre observation de G^* et l'information a priori. Puisque le logarithme ne change pas la position des extrema d'une fonction $f(x)$ dont l'image est $[0,1]$ (le logarithme est une fonction monotone et croissante), on préférera généralement l'estimateur équivalent

$$\hat{\boldsymbol{\theta}}^{\mathrm{MAP}} = \mathrm{argmax}_{\boldsymbol{\theta}}\big[\log P(\boldsymbol{\theta}|G^*)\big] = \mathrm{argmax}_{\boldsymbol{\theta}}\left[\log P(G^*|\boldsymbol{\theta}) + \log P(\boldsymbol{\theta})\right] \tag{1.34}$$

qui ne contient pas de référence à l'évidence[27], et qui explicite la relation additive entre l'information a priori et la vraisemblance, sous maximisation. Un avantage supplémentaire de cet estimateur logarithmique est la robustesse numérique de $\log P(\boldsymbol{\theta}|G^*)$ lorsque le support de $P(\boldsymbol{\theta}|G^*)$ est grand (si le support est exponentiellement grand, le logarithme reste de taille polynomiale).

Une autre stratégie d'estimation consiste à minimiser l'erreur attendue de l'estimateur [252]. Pour des paramètres continus, cette erreur peut s'écrire dans la forme quadratique $|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*|^2$, où $\boldsymbol{\theta}^*$ représente les paramètres de référence à inférer. Puisqu'on ne connaît pas $\boldsymbol{\theta}^*$, on utilise toute l'information disponible et on suppose $\boldsymbol{\theta}^* \sim P(\boldsymbol{\theta}|G^*)$, i.e., que les paramètres sont tirés de la distribution a posteriori. L'erreur quadratique moyennée sur la distribution a posteriori est alors

$$\phi(\hat{\boldsymbol{\theta}}) = \int P(\boldsymbol{\theta}|G^*)|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}|^2 d\boldsymbol{\theta}\,, \tag{1.35}$$

qu'on cherche à minimiser. Procédant par dérivation et utilisant la normalisation de la distribution a posteriori, on trouve que les estimateurs $\hat{\boldsymbol{\theta}}$ qui minimisent l'erreur en moyenne (les estimateurs MMSE, de *minimum mean-squared error*) sont

$$\hat{\boldsymbol{\theta}}^{\mathrm{MMSE}} = \int \boldsymbol{\theta}\, P(\boldsymbol{\theta}|G^*)\, d\boldsymbol{\theta} = \langle \boldsymbol{\theta} \rangle_{P(\boldsymbol{\theta}|G^*)}\,, \tag{1.36}$$

[27] Celle-ci n'intervient pas dans le calcul des probabilités a posteriori *relatives*. Elle peut donc être ignorée dans tous les calculs de maximisation. La formulation logarithmique est plus explicite en ce sens.

i.e., les premiers moments de la distribution a posteriori.

Pour des paramètres discrets, on peut se permettre d'utiliser une mesure d'erreur plus draconienne, i.e., de compter le nombre de paramètres inférés *exactement* [252]. Supposant qu'on doit estimer q paramètres, cette erreur s'écrit comme $q - \sum_{r=1}^{q} \mathbb{I}(\hat{\theta}_r = \theta_r^*)$, où $\boldsymbol{\theta}^*$ représente encore une fois les paramètres de référence, inconnus. Utilisant à nouveau $\boldsymbol{\theta}^* \sim P(\boldsymbol{\theta}|G^*)$ par manque d'information additionnelle, on attribue la probabilité

$$P(\theta_r|G^*) = \sum_{\theta_1} \dots \sum_{\theta_{r-1}} \sum_{\theta_{r+1}} \dots \sum_{\theta_q} P(\boldsymbol{\theta}|G^*) \tag{1.37}$$

à θ_r, par marginalisation de la distribution complète $P(\boldsymbol{\theta}|G^*)$. La méthode qui permettra de réduire l'erreur $q - \sum_{r=1}^{q} \mathbb{I}(\hat{\theta}_r = \theta_r^*)$ est évidemment celle qui maximise nos chances d'estimer correctement chaque paramètre, individuellement. On obtient ainsi les estimateurs de recouvrement maximal moyen (*maximum mean overlap*, MMO) :

$$\hat{\theta}_r^{\text{MMO}} = \text{argmax}_{\theta_r} \; P(\theta_r|G^*) \; , \tag{1.38}$$

i.e., les estimateurs qui assignent $\hat{\theta}_r$ au maximum de la distribution *marginale* a posteriori, pour tous les $r = 1, \dots, q$.

Estimation de paramètres : exemples

Considérons quelques exemples afin de concrétiser l'utilisation de ces estimateurs.

Erdős–Rényi. Le modèle ER permet d'estimer la densité d'un réseau rigoureusement. Bien que l'approche analytique soit dans ce cas quasiment tautologique, sa simplicité permet de bien illustrer l'estimation de paramètres sur réseaux. On se rappellera que la probabilité que le modèle ER génère un graphe donné $G(V, E)$ de N noeuds et M liens est [voir Eq. (1.6)]

$$P(G^*|\rho) = (1-\rho)^{\binom{N}{2} - M} \rho^M \; ,$$

où on a explicité le conditionnement sur ρ, maintenant considéré comme une variable aléatoire. Ce conditionnement nous donne directement la vraisemblance. On doit maintenant trouver la probabilité a priori. Pour simplifier le calcul, on choisit une distribution Beta(α, β), la distribution *conjuguée* de $P(G^*|\rho)$ [209]. Sa densité de probabilité est définie comme

$$P(\rho) = \frac{\rho^{\alpha-1}(1-\rho)^{\beta-1}}{B(\alpha,\beta)}, \qquad B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}, \qquad \alpha > 0, \beta > 0 \, . \tag{1.39}$$

Dans un contexte réseau, le paramètre $\alpha - 1$ est le nombre de liens attendu a priori, et le paramètre $\beta - 1$ est le nombre de non-liens a priori. Il est donc clair que si on ne connaît rien sur G^* a priori, on doit avoir $\alpha = \beta$ par symétrie, si on ne veut pas biaiser l'estimation.

La distribution a priori étant conjuguée à la vraisemblance, on trouve facilement que

$$P(\rho|G^*) \propto P(G^*|\rho)P(\rho) = (1-\rho)^{\binom{N}{2} - M + \beta - 1} \rho^{M + \alpha - 1} \; , \tag{1.40}$$

ce qui implique que la distribution a posteriori est également une distribution Beta

$$P(\rho|G^*) = \text{Beta}\left(M+\alpha, \binom{N}{2} - M + \beta\right) , \tag{1.41}$$

par normalisation. Or, les paramètres ont été mis à jour et contiennent maintenant de l'information sur le réseau. Puisque ρ est un paramètre continu, on peut trouver l'estimateur de maximum de vraisemblance par dérivation du logarithme de la distribution a posteriori. Pour x suivant une loi Beta(α, β), l'optimum est donné par

$$\left.\frac{\partial \log P(x)}{\partial x}\right|_{x=x*} = \frac{\alpha-1}{x^*} - \frac{\beta-1}{1-x^*} = 0 \quad \Longrightarrow \quad x^* = \frac{\alpha-1}{\alpha+\beta-2} \tag{1.42}$$

de sorte que

$$\hat{\rho}^{\text{MAP}} = \frac{M+\alpha-1}{\binom{N}{2}+\alpha+\beta-2} . \tag{1.43}$$

L'estimateur qui minimise l'erreur quadratique est quant à lui donné par la moyenne de la distribution a posteriori, voir Eq. (1.36). Puisque la moyenne d'une distribution Beta de paramètres (α, β) est donnée par $\alpha/(\alpha+\beta)$, on trouve

$$\hat{\rho}^{\text{MMSE}} = \frac{M+\alpha}{\binom{N}{2}+\alpha+\beta} . \tag{1.44}$$

Dans la limite des grands réseaux où M diverge, N doit diverger par nécessité, de sorte que le poids de l'hypothèse initiale sur ρ^* diminue, menant ultimement a

$$\hat{\rho}^{\text{MAP}} \sim \hat{\rho}^{\text{MMSE}} \sim \frac{2M}{N(N-1)} , \tag{1.45}$$

indépendamment du choix de la loi a priori[28]. On obtient le même résultat pour N petit si on utilise la loi uniforme ($\alpha = \beta = 1$), comme distribution a priori sur ρ ; c'est l'équivalent d'ignorer $P(\rho)$, ce qui mène à un calcul directement sur la vraisemblance $P(\rho|G^*) \propto P(G^*|\rho)$, i.e., à une approche non bayésienne.

Modèles des configurations. L'estimation est plus complexe si on considère un modèle moins trivial, comme le modèle CM canonique. On se rappellera que dans cette version du CM, les paramètres cachés sont les degrés attendus $\boldsymbol{\kappa}$. S'inspirant de l'analyse du modèle ER, on peut se douter qu'on obtiendra $\hat{\kappa}_i \to k_i$ pour ces paramètres. Vérifions cette intuition à l'aide d'un calcul explicite. On débute avec la vraisemblance introduite à l'Eq. (1.9) :

$$P(G^*|\boldsymbol{\kappa}) = \prod_{i<j} \frac{(\kappa_i\kappa_j/2\mu)^{a_{ij}} e^{-\kappa_i\kappa_j/2\mu}}{a_{ij}!} , \tag{1.46}$$

[28] Le nombre de liens et non-liens observés joue le rôle de la taille de l'échantillon. On a convergence de l'estimateur dans la limite des grands réseaux, même si on observe qu'un seul réseau.

où $2\mu := \sum_i \kappa_i$. Une réorganisation similaire à celle utilisée dans le cas du modèle ER facilite le traitement mathématique. En effet, on observe que

$$\prod_{i<j} e^{-\kappa_i\kappa_j/2\mu} = \exp\left[-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\kappa_i\kappa_j}{2\mu} + \frac{1}{2}\sum_{i=1}^{N}\frac{\kappa_i^2}{2\mu}\right] = \exp\left[\frac{1}{2}\sum_{i=1}^{N}\frac{\kappa_i^2}{2\mu} - \mu\right] = \prod_{i=1}^{N}\exp\left[\frac{\kappa_i^2}{4\mu} - \frac{\mu}{N}\right]$$

et que $a_{ij}! = 1$ pour un graphe simple, de sorte qu'on peut écrire la vraisemblance comme un produit sur les noeuds plutôt que sur les liens :

$$P(G^*|\boldsymbol{\kappa}) = \prod_{i=1}^{N}\left(\frac{\kappa_i^2}{2\mu}\right)^{\frac{1}{2}\sum_j a_{ij}} \exp\left[\frac{\kappa_i^2}{4\mu} - \frac{\mu}{N}\right] . \tag{1.47}$$

On suppose une distribution a priori de Dirichlet sur $\boldsymbol{\kappa}/2\mu$, car cette distribution force une normalisation des variables aléatoires, i.e., $\sum_i \kappa_i/2\mu = 1$. Sa densité de probabilité est donnée par

$$P(\boldsymbol{\kappa};\boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})}\prod_{i=1}^{N}\left(\frac{\kappa_i}{2\mu}\right)^{\alpha_i - 1} \qquad \alpha_i > 0 \; \forall \; i , \tag{1.48}$$

où les $\boldsymbol{\alpha}$ contrôlent la concentration de la distribution et où $B(\boldsymbol{\alpha})$ est une normalisation. On cherche d'abord l'estimateur MAP. Les paramètres cachés $\boldsymbol{\kappa}$ sont des variables continues, ce qui implique qu'on peut trouver l'optimum par dérivation. Pour simplifier le calcul, on considère μ comme une constante, et on introduit une équation de contrainte additionnelle $\sum \kappa_i = 2\mu$, à l'aide du multiplicateur de Lagrange λ[29]. La dérivée de la distribution a posteriori est donc proportionnelle à

$$\frac{\partial \log P(G^*|\boldsymbol{\kappa})P(\boldsymbol{\kappa}) - \lambda(\sum\kappa_i - 2\mu)}{\partial k_r} = \sum_{i=1}^{N}\left[\frac{k_i}{\kappa_i}\delta_{ir} + \frac{\kappa_i}{2\mu}\right] + \sum_{i=1}^{N}\left[\frac{\alpha_i - 1}{\kappa_i}\delta_{ir}\right] - \lambda , \tag{1.49}$$

où la première somme provient de la vraisemblance, la deuxième de la distribution a priori, et où λ est dû à l'équation de contrainte. Utilisant la normalisation pour simplifier le ratio $\sum\kappa_i/2\mu = 1$ dans la première somme, on obtient que les points critiques doivent solutionner

$$k_r + \alpha_r - 1 = \hat{\kappa}_r(\lambda - 1) . \tag{1.50}$$

Finalement, utilisant l'équation de contraintes à nouveau pour trouver la valeur de λ, on obtient

$$\hat{\kappa}_r^{\text{MAP}} = \frac{k_r + \alpha_r - 1}{\sum_r(\alpha_r - 1)/2M + 1} , \tag{1.51}$$

après avoir choisi la normalisation $\mu = M$, imposée par les données.

Pour une distribution a priori uniforme, obtenue en choisissant $\alpha_i = 1$ pour $i = 1, \ldots, N$, on trouve sans surprise que l'estimateur (1.51) donne $\hat{\kappa}_r^{\text{MAP}} = k_r$. Mais lorsque la distribution a priori est non-uniforme, l'estimé peut varier grandement, selon le régime de densité observé.

[29] De façon équivalente, on peut voir μ comme une fonction de $\boldsymbol{\kappa}$, mais le calcul est moins simple et moins élégant.

En effet, si les éléments de $\boldsymbol{\alpha}$ sont finis [30], et le graphe est uniformément dense (tous les degrés divergent), alors k_r domine $\alpha_r - 1$ et $\sum_r(\alpha_r - 1)/2M \to 0$ dans la limite $N \to \infty$, de sorte qu'on obtient encore $\hat{\kappa}_r^{\text{MAP}} \to k_r$, indépendamment de la distribution a priori. Mais si le graphe est uniformément creux et que k_r est une constante dans la limite $N \to \infty$, alors $M = O(N)$ et *aucun* terme domine dans l'Eq. (1.51). Il n'y a tout simplement pas assez d'observations pour effacer la contribution de l'information a priori. Finalement, si seuls quelques degrés divergent, alors on a $\hat{\kappa}_r^{\text{MAP}} \to k_r$ pour ces noeuds, et $\hat{\kappa}_r \to k_r + \alpha_r - 1$ pour les autres.

Le calcul de l'estimateur κ_r^{MMSE} est plus complexe car la distribution de Dirichlet n'est pas conjuguée à la vraisemblance (1.47). Ce cas d'étude illustre donc bien que l'inférence peut devenir complexe : même pour le cas relativement simple du CM canonique, la forme fermée de $\hat{\boldsymbol{\kappa}}^{\text{MMSE}}$ ne peut être obtenue sans le calcul d'une intégrale compliquée. On pourrait contourner ce problème en recourant à une évaluation numérique par échantillonnage Monte-Carlo (voir Sec. 1.3 de ce chapitre pour des exemples). Mais on se contentera pour l'instant d'un seul estimateur, $\boldsymbol{\kappa}^{\text{MAP}}$.

ERGMs. La portée des méthodes analytiques est encore plus faible pour des ERGMs parfaitement généraux—l'estimation de paramètres par méthodes analytiques est dans ce cas tout simplement impossible. En effet, la vraisemblance du réseau G^* introduite en (1.20) s'écrit comme

$$P(G^*|\boldsymbol{\beta}) = \frac{e^{\sum_i \beta_i \theta_i(G^*)}}{Z(\boldsymbol{\beta})} \qquad \text{où} \qquad Z(\boldsymbol{\beta}) = \sum_{\{G\}} e^{\sum_i \beta_i \theta_i(G)} \,, \tag{1.52}$$

lorsqu'on explicite le conditionnement sur les paramètres $\boldsymbol{\beta}$. Cette vraisemblance est conjuguée à une famille exponentielle de distributions [58, Eq. (2.3)], qu'on peut écrire sous la forme simplifiée

$$P(\boldsymbol{\beta}) = \frac{e^{-\sum_i \beta_i \eta_i}}{\zeta(\boldsymbol{\eta})} \qquad \text{où} \qquad \zeta(\boldsymbol{\eta}) = \int d\beta_1 \dots d\beta_q \; e^{-\sum_{i=1}^{q} \beta_i \eta_i} \,, \tag{1.53}$$

d'hyperparamètres $\boldsymbol{\eta}$. Suivant la même démarche que dans les exemples précédents, on vérifie que l'estimateur $\boldsymbol{\beta}_{\boldsymbol{r}}^{\text{MAP}}$ est donné par la solution de

$$\theta_r(G) - \sum_{\{G\}} \theta_r(G) \frac{e^{\sum_i \hat{\beta}_i \theta_i(G))}}{Z(\hat{\boldsymbol{\beta}})} - \eta_r = 0 \,, \tag{1.54}$$

obtenu par dérivation de la distribution a posteriori. Or, il s'agit tout simplement de la définition implicite

$$\theta_r(G^*) = \langle \theta_r(\boldsymbol{\beta}^{\widehat{\text{MAP}}}) \rangle \,, \tag{1.55}$$

à laquelle on a ajouté un paramètre η_r, signature de l'information a priori. Ainsi, en d'autres mots, les paramètres qui maximisent la distribution a posteriori sont ceux qui fixent les moyennes de l'ensemble aux valeurs observées dans G^*, avec une contribution négligeable

[30] Toujours le cas, sinon on donne un poids infini à l'hypothèse a priori, ignorant ainsi les données.

de l'information a priori lorsqu'on choisit la distribution uniforme ($\eta_i \to 0$ pour $i = 1, \dots, q$). Cette forme est, en pratique, parfaitement inutile, car la fonction de partition $Z(\boldsymbol{\beta})$ est trop difficile à calculer pour la plupart des choix de statistique $\boldsymbol{\theta}$ arbitraire, surtout si N est grand [34, 74]. On doit donc à nouveau avoir recours à des méthodes numériques pour solutionner (1.55). L'estimateur MMSE est tout aussi difficile à évaluer, pour des raisons similaires.

Modèle stochastique par blocs . Pour clore cette section sur l'estimation de paramètres, on retourne au modèle stochastique par blocs (canonique), lequel est différent des exemples précédents car il incorpore deux ensembles de paramètres distincts en un seul modèle : les assignations à des blocs $\boldsymbol{\sigma}$ et une matrice de densité $\boldsymbol{P}$. On se rappellera que sa vraisemblance est donnée par

$$P(G^*|\boldsymbol{\theta}) = \prod_{r\leq s}(1-p_{rs})^{m_{rs}^{\max}-m_{rs}} p_{rs}^{m_{rs}}\,, \qquad m_{rs}^{\max} = \begin{cases} n_r n_s & \text{si } r \neq s, \\ \binom{n_r}{2} & \text{sinon,} \end{cases}$$

où est m_{rs} est le nombre de liens entre les noeuds assignés au bloc $\sigma(i) = r$ et $\sigma(j) = s$, et où $\boldsymbol{\theta}$ dénote les paramètres, collectivement. On note que la variable $\boldsymbol{\sigma}$ est discrète, et que son support est exponentiellement grand en N (jusqu'à q^N possibilités, si les groupes vides sont permis), tandis que la variable $\boldsymbol{P}$ est continue, et que son support est un hypercube de dimension polynomiale en q. Ces différences suggèrent un traitement distinct pour les deux ensembles de paramètres. On explicite cette différence en écrivant la formule de Bayes comme

$$P(\boldsymbol{\sigma}, \boldsymbol{P}|G^*) = \frac{P(G^*|\boldsymbol{\sigma}, \boldsymbol{P})P(\boldsymbol{P}, \boldsymbol{\sigma})}{P(G^*)}\,. \tag{1.56}$$

On peut se douter d'emblée que $\boldsymbol{\sigma}$ causera plus de problèmes que $\boldsymbol{P}$, puisque les méthodes de l'analyse ne peuvent pas nous aider (la variable est discrète), et qu'il y a énormément de possibilités à tenir en compte. Il faudra en fait développer des outils numériques pour manipuler cette portion de la distribution a posteriori (c.f., Sec. 1.3 qui suit). Toutefois, avant de se lancer dans un traitement numérique, on souligne quand même deux exemples de calcul d'estimation qui peuvent être faits analytiquement, dans l'esprit des calculs de la présente section. Ces calculs interviendront dans les outils numériques de la section 1.3.

Premier calcul. Si on suppose $\boldsymbol{\sigma}$ connu, la formule de Bayes se réduit à

$$P(\boldsymbol{P}|G^*) = \frac{P(G^*|\boldsymbol{P})P(\boldsymbol{P})}{P(G^*)}\,.$$

Dans ce cas particulier, on peut obtenir facilement des estimateurs à l'aide de la méthode utilisée pour le modèle ER. En effet, supposant une distribution Beta multivariée sur $\boldsymbol{P}$, on obtient

$$P(\boldsymbol{P}|G^*) \propto P(G^*|\boldsymbol{P})P(\boldsymbol{P}) = \prod_{r\leq s}(1-p_{rs})^{m_{rs}^{\max}-m_{rs}+\beta_{rs}-1} p_{rs}^{m_{rs}+\alpha_{rs}-1}\,. \tag{1.57}$$

Puisque les termes du produit sont tous séparés, on trouve directement que

$$\hat{p}_{rs}^{\text{MAP}} = \frac{m_{rs}+\alpha_{rs}-1}{m_{rs}^{\max}+\alpha_{rs}+\beta_{rs}-2}\,, \qquad \hat{p}_{rs}^{\text{MMSE}} = \frac{m_{rs}+\alpha_{rs}}{m_{rs}^{\max}+\alpha_{rs}+\beta_{rs}} \tag{1.58}$$

i.e., $\hat{p}_{rs} \approx m_{rs}/m_{rs}^{\max}$. Ce calcul nous apprend donc que pour une division en blocs fixe, les estimateurs optimaux de p_{rs} peuvent être calculés rapidement, en comptant le nombre de liens entre les blocs.

Deuxième calcul. Supposons qu'on cherche seulement $\boldsymbol{\sigma}$. Il convient dans ce cas d'intégrer le modèle, i.e., de calculer

$$P(\boldsymbol{\sigma}|G^*) = \frac{\int P(G^*|\boldsymbol{P},\boldsymbol{\sigma})P(\boldsymbol{P},\boldsymbol{\sigma})d\boldsymbol{P}}{P(G^*)}\,. \tag{1.59}$$

Puisque nous avons un contrôle total sur les distributions a priori, on peut simplifier l'inférence et se concentrer sur des distributions jointes indépendantes pour lesquelles $P(\boldsymbol{P},\boldsymbol{\sigma}) = P(\boldsymbol{P})P(\boldsymbol{\sigma})$, i.e.,

$$P(\boldsymbol{\sigma}|G^*) = \frac{P(\boldsymbol{\sigma})\int P(G^*|\boldsymbol{P},\boldsymbol{\sigma})P(\boldsymbol{P})d\boldsymbol{P}}{P(G^*)}\,.$$

Malgré cette simplification, la distribution a posteriori implique une intégrale compliquée, qu'on écrit comme

$$\begin{aligned}
I(\boldsymbol{\sigma}) &= \int_{[0,1]^{q(q+1)/2}} \exp\left\{\sum_{r\leq s}\left[(m_{rs}^{\max} - m_{rs})\log(1-p_{rs}) + m_{rs}\log p_{rs}\right]\right\} dp_{11}dp_{12}\dots dp_{qq} \\
&= \prod_{r\leq s}\int_0^1 dp_{rs}\,\exp\left\{m_{rs}^{\max}\left[\left(1-\frac{m_{rs}}{m_{rs}^{\max}}\right)\log(1-p_{rs}) + \frac{m_{rs}}{m_{rs}^{\max}}\log p_{rs}\right]\right\}\,.
\end{aligned}$$

L'intégrale peut être calculée à l'aide de la méthode du col. En effet, il est facile de vérifier par différenciation de l'argument $g(p_{rs})$ de la fonction exponentielle que la contribution maximale à l'intégrale se situe à $\widehat{p_{rs}} = m_{rs}/m_{rs}^{\max}$. On peut donc approximer l'intégrale par

$$I(\boldsymbol{\alpha}) \approx \prod_{r\leq s} e^{-m_{rs}^{\max}h(m_{rs}/m_{rs}^{\max})} = \prod_{r\leq s}\left(1-\frac{m_{rs}}{m_{rs}^{\max}}\right)^{m_{rs}^{\max}-m_{rs}}\left(\frac{m_{rs}}{m_{rs}^{\max}}\right)^{m_{rs}} \tag{1.60}$$

où $h = -x\log x - (1-x)\log(1-x)$ est l'entropie associée à une variable de Bernoulli de paramètre x. La dépendance sur $\boldsymbol{\alpha}$ se manifeste via le calcul de $\boldsymbol{m}^{\max}$ et $\boldsymbol{m}$. Ce calcul nous donne une distribution a posteriori indépendante de $\boldsymbol{P}$, équivalente à la distribution a posteriori du SBM microcanonique [voir Eq. (7) de la Ref. [190]].

1.3 Algorithmes et échantillonnage

> *I propose we leave math to the machines and go play outside.*
>
> —Calvin

À maintes reprises, on a rencontré des problèmes algorithmiques, jusqu'à maintenant escamotés : échantillonnage des modèles aléatoires (Sec. 1.1.2), calcul de la distribution des observables de modèles nuls (Sec. 1.2.1), estimation des paramètres de modèles aléatoires canoniques (Sec. 1.2.2). On conclut ce chapitre d'introduction en attaquant ces problèmes de front, à l'aide d'un outil très puissant : le principe de Monte-Carlo[31].

[31] L'organisation de cette dernière section est inspirée de la référence [11].

1.3.1 Principe de Monte-Carlo

Dans tous les problèmes énoncés ci-haut, on cherche à calculer des fonctions compliquées de variables aléatoires difficiles à décrire analytiquement. Le principe de Monte-Carlo (MC) permet d'y arriver numériquement. L'idée générale de la méthode MC est d'échantillonner des éléments $\{x^{(i)}\}_{i=1}^{n}$ indépendants et identiquement distribués (i.i.d.) de la densité $p(x)$ de X, puis de les utiliser pour approximer la densité $p(x)$ par

$$p_n(x) = \frac{1}{n}\sum_{i=1}^{n}\delta(x - x^{(i)})\ , \tag{1.61}$$

où $\delta(x)$ est le delta de Dirac dans le cas continu (et où on peut le remplacer par le delta de Kronecker dans le cas discret). On peut alors se servir de la distribution empirique approximative $p_n(x)$ pour calculer, par exemple : les moments, la distribution ou les fonctions de X. En effet, on vérifie en substituant $p(x)$ par $p_n(x)$ que :

X continue	X discrète
$\langle f(X)\rangle = \int_{\mathcal{X}} f(x)p(x)dx \approx \frac{1}{n}\sum_{i=1}^{n} f(x^{(i)})$	$\langle f(X)\rangle = \sum_{x\in\mathcal{X}} f(x)p(x) \approx \frac{1}{n}\sum_{i=1}^{n} f(x^{(i)})$
$p(x\in A) = \int_A p(x)dx \approx \frac{\sum_i \mathbb{I}[x^{(i)}\in A]}{n}$	$p(X=x) \approx p_n(x)$

Dans tous les cas, la convergence de l'approximation est garantie dans la limite $n \to \infty$ par la loi des grands nombres.

Le principe de Monte-Carlo est, sans aucun doute, très puissant. Il ne mène toutefois pas à une prescription *automatique* pour le calcul des fonctions de X, car la construction de l'échantillon i.i.d. $\{x^{(i)}\}_{i}^{n}$ est hautement non triviale. Tout l'art de l'application du principe MC réside dans la création d'algorithmes permettant de générer ces échantillons efficacement. On en survole ici quelques-uns qui seront revisités plus tard dans la thèse : la méthode de rejet (chapitre 5), la méthode de l'échantillonnage préférentiel (chapitre 4), et les méthodes de Monte-Carlo par chaînes de Markov (chapitres 2–3 et 5).

1.3.2 Méthode de rejet

Stories only happen to those who are able to tell them.
—Paul Auster

Grossièrement, l'idée de la méthode de rejet est d'utiliser une variable aléatoire Y simple à échantillonner (la distribution uniforme, par exemple) pour construire un échantillon aléatoire d'une variable X plus compliquée à manipuler. On dit qu'on *propose* des échantillons aléatoires, qu'on *accepte* sous certaines conditions, choisies de façon à ce que l'échantillon retenu soit distribué i.i.d. selon la densité cible. Il est clair que pour que la méthode fonctionne, il faut que la distribution de proposition ait un support $\mathcal{Y}$ qui contient le support $\mathcal{X}$ de la densité cible, i.e., $\mathcal{X} \subset \mathcal{Y}$. Lorsque c'est le cas, il existe une constante finie λ tel que

$p(x) \leq \lambda q(x)$ pour tout $x \in \mathcal{X}$, où $p(x)$ est la densité de X et $q(x)$ est la densité de Y. On dit alors que la densité $\lambda q(x)$ est *l'enveloppe* de $p(x)$ parce qu'elle la borne supérieurement pour tout x. Toute densité qui enveloppe la densité cible peut être utilisée comme distribution de proposition pour la méthode de rejet. La méthode de rejet consiste alors à

1. échantillonner $x \sim Y$;
2. accepter l'échantillon avec probabilité $p(x)/\lambda q(x)$, ou retourner à l'étape 1 avec probabilité complémentaire.

À chaque itération de la boucle, la probabilité d'accepter un échantillon est donnée par

$$\int_{\mathcal{Y}} q(x)\frac{p(x)}{\lambda q(x)}dx = \frac{1}{\lambda}\int_{\mathcal{Y}} p(x)dx = \frac{1}{\lambda}\int_{\mathcal{X}} p(x)dx = \frac{1}{\lambda}\,. \tag{1.62}$$

Puisque chaque boucle de l'algorithme mène à un échantillon indépendant, il faudra en moyenne λ boucles avant de rencontrer un échantillon accepté[32]. La probabilité d'accepter un échantillon dans l'intervalle A est donc, en moyenne,

$$\lambda \int_{A} q(x)\frac{p(x)}{\lambda q(x)}dx = \int_{A} p(x)dx\,, \tag{1.63}$$

ce qui démontre que l'algorithme de rejet fonctionne tel qu'escompté.

Méthode de rejet : Exemple

Afin d'illustrer cette méthode dans le contexte de l'inférence réseau, considérons le CM microcanonique sans boucles (v_i,v_i) et sans lien multiple, i.e., la distribution uniforme sur $\Omega_N(\boldsymbol{k})$, l'ensemble des réseaux simples de N noeuds, dont la séquence de degrés est $\boldsymbol{k}$. Échantillonner ce modèle n'est pas évident [79]. Or, il existe un algorithme efficace et simple, dû à B. Bollobás [26], qui permet d'échantillonner uniformément un ensemble $\Psi_N(\boldsymbol{k}) \supset \Omega_N(\boldsymbol{k})$ plus grand, soit celui de tous les réseaux dont la séquence de degrés est $\boldsymbol{k}$, sans la spécification additionnelle que ces réseaux soient simples. L'algorithme se décrit comme suit :

1. générer une liste d'entiers L de longueur $2M = \sum_{i=1}^{N} k_i$, où i apparaît k_i fois ;
2. choisir une permutation aléatoire de L, uniformément parmi toutes les permutations possibles ;
3. lire la liste par paires (l_1,l_2), (l_3,l_4), …, (l_{2M-1},l_{2M}) et l'interpréter comme une liste de liens pour un graphe G.

Lorsqu'on interprète la liste L comme une liste de demi-liens, il devient évident que G aura la séquence de degrés $\boldsymbol{k}$ par construction, et donc que l'étape 2 engendre une distribution uniforme sur tous les graphes de séquence $\boldsymbol{k}$. Cependant, la distribution est uniforme sur $\Psi_N(\boldsymbol{k})$ plutôt que $\Omega_N(\boldsymbol{k})$, parce que l'algorithme peut mener à des liens multiples et des boucles—rien n'interdit les paires $(l_t,l_{t+1}) = (v_i,v_i)$, ou encore qu'une paire particulière (v_i,v_j) n'apparaisse à deux reprises dans L.

[32] Le nombre d'essais nécessaires suit une distribution géométrique de paramètre $1/\lambda$.

Pour échantillonner l'ensemble $\Omega_N(\boldsymbol{k})$ de graphes *simples* associé à $\boldsymbol{k}$, une option consiste à exploiter la propriété $\Psi_N(\boldsymbol{k}) \supset \Omega_N(\boldsymbol{k})$ et à utiliser l'algorithme de Bollobás comme distribution de proposition pour la méthode de rejet. La probabilité d'acceptation d'un échantillon est donnée par $p(G)/\lambda q(G)$, i.e.,

$$\text{Prob(accepter } G) = \begin{cases} \frac{1/|\Omega_N(\boldsymbol{k})|}{\lambda/|\Psi_N(\boldsymbol{k})|} & \text{si } G \in \Omega_N(\boldsymbol{k}) \\ 0 & \text{sinon,} \end{cases} \tag{1.64}$$

Il est facile de voir qu'on peut maximiser le taux d'acceptation global en choisissant

$$\lambda = \frac{|\Psi_N(\boldsymbol{k})|}{|\Omega_N(\boldsymbol{k})|}\,, \tag{1.65}$$

de façon à toujours accepter un réseau proposé quand il est dans $\Omega_N(\boldsymbol{k})$. La constante λ étant spécifiée, l'algorithme est complet et on peut l'utiliser, voir Fig. 1.3.

Notons que la valeur λ apparaissant à l'Eq. (1.65) permet de formaliser l'intuition qu'une trop grande différence de taille mène à un algorithme inefficace [33], puisqu'il faut en moyenne λ essais avant d'obtenir un échantillon acceptable. Idéalement, on aurait $|\Psi_N(\boldsymbol{k})| = O(|\Omega_N(\boldsymbol{k})|$, i.e., un nombre constant d'essais. Mais ce n'est normalement pas le cas ; l'ensemble de graphes simples est généralement beaucoup plus petit que l'ensemble sans contraintes correspondant. Par exemple, $|\Psi_N(\boldsymbol{k})|/|\Omega_N(\boldsymbol{k})|$ est de l'ordre de 10^4 pour la séquence $\boldsymbol{k}$ du réseau de dauphins de la section 1.2.1 (voir Fig. 1.3). On préférera donc généralement d'autres méthodes pour ce genre de problèmes d'échantillonnage, par exemple les chaînes de Markov Monte-Carlo (c.f. Sec. 1.3.4 et chapitre 5). La méthode de rejet reste utile pour des petits réseaux, ou pour d'autres distributions.

1.3.3 Échantillonnage préférentiel

> *When men begin to do luxurious and idle work,*
> *it is inevitable that a few do all the exchange work.*
> —Henry David Thoreau

L'échantillonnage préférentiel ressemble superficiellement à la méthode de rejet, parce qu'il repose aussi sur l'utilisation d'une distribution de proposition $q(x)$ pour calculer des intégrales de $p(x)$. L'objectif n'est toutefois plus de générer des échantillons i.i.d. de $p(x)$, la distribution cible, mais plutôt de générer des échantillons *importants*, i.e., qui comptent pour beaucoup dans l'évaluation des intégrales [11]. Un choix judicieux de distribution $q(x)$ permet d'accélérer la convergence en réduisant la variance des estimateurs [209].

[33] Voir chapitre 5, où on discute d'un algorithme de rejet pour une généralisation du CM, et où on utilise cet argument pour justifier le rejet... de la méthode de rejet !

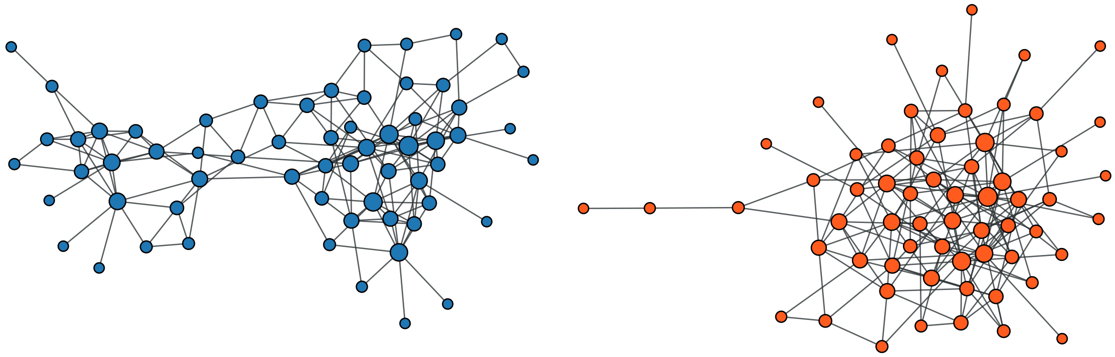


FIGURE 1.3 – (gauche) Réseau des interactions sociales d'un groupe de dauphins, reproduit de la Figure 1.1, sans séparation en communautés. (droite) Réseau aléatoire avec la même séquence de degrés $\boldsymbol{k}$ que le réseau de dauphins, généré à l'aide de la méthode de rejet (après 7495 rejets).

Le point de départ est la réorganisation

$$I(f) = \int_{\mathcal{X}} f(x)p(x)dx = \int_{\mathcal{Y}} f(x)q(x)w(x)dx = \langle f(x)w(x)\rangle_q \tag{1.66}$$

où $w(x) = p(x)/q(x)$ est appelé le poids de x. Le principe de Monte-Carlo nous dit qu'on peut approximer cette intégrale à l'aide d'un échantillon $\{x^{(i)}\}_{i=1\ldots n}$ i.i.d. tiré de la distribution de densité $q(x)$, via

$$\hat{I}_n(f) = \frac{1}{n}\sum_{i=1}^{n} f(x^{(i)})w(x^{(i)})\ . \tag{1.67}$$

La loi des grands nombres garantit encore que $\hat{I}_n(f) \to I(f)$ dans la limite $n \to \infty$. La différence importante avec le principe de Monte-Carlo pur est que la distribution $q(x)$ permet maintenant de contrôler la vitesse de convergence. En effet, la variance de l'estimateur est fonction de q :

$$\mathrm{Var}_q[f(x)w(x)] := \langle f(x)^2 w(x)^2\rangle_q - \langle f(x)w(x)\rangle_q^2 = \langle f(x)^2 w(x)^2\rangle_q - I(f)^2\ . \tag{1.68}$$

On peut la minimiser en minimisant $\langle f(x)^2 w(x)^2\rangle_q$ via $q(x)$, puisque le choix de distribution $q(x)$ n'affecte pas le terme négatif $-I(f)^2$. Un calcul explicite montre qu'on peut atteindre le minimum théorique de la variance [$\mathrm{Var}_q(f(x)w(x)) = 0$] en choisissant $q = q^*$ comme

$$q^* = \frac{|f(x)|p(x)}{\int_{\mathcal{Y}} |f(x)|p(x)dx} \quad\Longrightarrow\quad \langle f(x)^2 w(x)^2\rangle_q = \langle f(x)w(x)\rangle_q^2\,. \tag{1.69}$$

Or, cette distribution q^* est tellement artificielle qu'elle est pratiquement impossible à échantillonner ; on doit normalement se contenter de distributions moins efficaces en termes de réduction de variance, mais plus pratique numériquement. La distribution q^* reste toutefois utile parce qu'elle nous enseigne que pour minimiser la variance de $\hat{I}_n(f)$, il faut concentrer $q(x)$ sur les points du support où $p(x)|f(x)|$ est grand.

Échantillonnage préférentiel : Exemple

À titre d'exemple d'application de l'échantillonnage préférentiel, on considère une version simplifiée du problème étudié dans le chapitre 4, et on profite de l'occasion pour revisiter le modèle d'attachement préférentiel introduit plus tôt (c.f., Sec. 1.1.2). On se rappellera que dans le modèle PA classique, un réseau est généré séquentiellement à partir d'un réseau initial contenant un seul noeud isolé. À chaque temps discret $t = 1, ..., T$, on forme un nouveau lien en attachant un nouveau noeud à un noeud déjà existant, sélectionné avec probabilité

$$p_i(t) = \frac{k_i^\gamma(t-1)}{\sum_i k_i^\gamma(t-1)} \,. \tag{1.70}$$

Pour simplifier l'exemple, on se concentre sur les deux cas spéciaux donnés par $\gamma = 0$ ou $\gamma = 1$ (attachement aléatoire et préférentiel, respectivement).

Pour poser le problème d'échantillonnage, on suppose qu'on a un réseau G à disposition, et qu'on pense qu'il est bien modélisé par le processus d'attachement préférentiel. On cherche à calculer la probabilité $P(\tau_e = 1)$ que le lien e ait été le premier lien du réseau. Cette probabilité est par définition

$$P(\tau_e = 1) := \text{Prob}[e \text{ est le premier lien}] = \sum_{X \in \pi(E(G))} \mathbb{I}[e \text{ est le premier lien dans } X] P(X|G) \tag{1.71}$$

où $P(X|G)$ est la probabilité a posteriori de l'ordonnement X (l'ordre d'apparition des liens), et où $\pi(E(G))$ est l'ensemble des permutations des liens de G, i.e., toutes les façons d'ordonner les liens de G. En principe, l'Eq. (1.71) détermine complètement la probabilité recherchée. Mais en pratique, puisque $|\pi(E(G))| = M!$, il est impossible de calculer $P(\tau_e = 1)$ explicitement, pour des réseaux de plus de quelques noeuds. Il faut trouver une solution approximative.

Une première option consiste à recourir au principe de Monte-Carlo pour estimer la probabilité $P(\tau_e = 1)$ à l'aide de

$$P(\tau_e = 1) \approx \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}[e \text{ est le premier lien dans } x^{(i)}] \tag{1.72}$$

où $x^{(i)}$ est tiré de la distribution a posteriori $P(X|G)$. Pour ce faire, il faut d'abord caractériser $P(X|G)$. On peut montrer que pour les valeurs particulières de l'exposant du noyau d'attachement considérées ($\gamma = 0$ et $\gamma = 1$), cette distribution est donnée par[34]

$$P(X|G) \propto P(G|X)P(X) = \frac{\mathbb{I}[X \in \Psi(G)]}{|\Psi(G)|} \,, \tag{1.73}$$

[34] L'intuition derrière ce résultat est que $p_i(t) = 1/m(t)$ où $m(t)$ est une fonction linéaire de t pour tout (i,t) lorsque $\gamma \in \{0,1\}$. Le résultat final suit par normalisation de la distribution. Voir chapitre 4 pour un détaillé.

où $\Psi(G) \subset \pi(E(G))$ est l'ensemble des ordonnements pouvant être produits par le mécanisme d'attachement, sachant qu'on a observé G[35]. Ainsi, la distribution a posteriori du modèle d'attachement est uniforme sur tous les $X \in \Psi(G)$. On peut donc évaluer l'Eq. (1.72) en générant des permutations uniformes de $E(G)$ et en rejetant celles qui ne sont pas dans $\Psi(G)$ [ou de façon équivalente : en notant que $P(X|G) = 0$ si $X \notin \Psi(G)$]. Or, si on se contente d'utiliser la méthode de rejet, l'estimateur MC convergera trop lentement, puisque la plupart des échantillons seront ignorés ; l'espace $\pi(E(G))$ est *beaucoup* plus grand que le support de $P(X|G)$.

On peut remédier à la situation en utilisant l'échantillonnage préférentiel pour produire uniquement des ordonnements dans $\Psi(G)$. Une façon simple d'y arriver consiste à énumérer les liens de G au hasard, certes, mais en s'assurant de respecter la connectivité de G à tous les temps de l'ordonnement. Plus précisément, après avoir choisi un premier lien $e_1 \in E(G)$ au hasard, on choisi un second lien e_2 uniformément dans l'ensemble $\Omega_2(G,X)$ des liens non énumérés partageant un noeud avec e_1, puis on répète ensuite la procédure avec $\Omega_3(G)$, l'ensemble de liens non énumérés partageant un noeud avec e_0 ou e_1, etc. Il est facile de se convaincre que cette procédure d'échantillonnage par avalanche (*snowball sampling*) peut générer tous les ordonnements $X \in \Psi(G)$ avec une probabilité non nulle, donnée par

$$Q(X) = \prod_{t=1}^{T} \frac{1}{|\Omega_t(G,X)|}\,, \tag{1.74}$$

où $\Omega_1(G) = E(G)$.

Utilisant alors l'Eq. (1.67) pour appliquer le principe de Monte-Carlo à l'échantillonnage préférentiel, on trouve que pour un échantillon $\{x^{(i)}\}_{i=1,\ldots,n}$ généré par échantillonnage en avalanche de G,

$$\begin{aligned} P(\tau_e = 1) &\approx \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}[e \text{ est le premier lien dans } x^{(i)}]\frac{P(X|G)}{Q(X)} \\ \implies \quad P(\tau_e = 1) &\propto \sum_{i=1}^{n} \mathbb{I}[e \text{ est le premier lien dans } x^{(i)}]\prod_{t=1}^{T}|\Omega_t(G,X)|\,, \end{aligned} \tag{1.75}$$

où on a ignoré la constante $|\Psi(G)|$ a priori inconnue. On obtient les probabilités désirées en faisant appel à la normalisation de la distribution $\sum_{e\in E(G)} P(\tau_e = 1) = 1$.

L'utilité de l'échantillonnage préférentiel est démontrée dans les figures 1.4 et 1.5, où on l'applique à des réseaux artificiels, générés à l'aide du PA d'exposant $\gamma = 0$ et $\gamma = 1$. On peut s'attendre à obtenir des probabilités $P(\tau_e = 1)$ sensées pour ces réseaux, puisqu'ils sont générés par le modèle $P(G)$ utilisé pour construire l'estimateur de $P(\tau_e = 1)$. C'est bien ce qui est observé, puisque dans le cas de l'exposant $\gamma = 0$ (Fig. 1.4), par exemple, l'échantillonnage

[35] L'ensemble $\Psi(G)$ exclut les ordonnements qui mènent à un réseau G déconnecté à au moins un point de l'énumération, e.g., $X = [(v_0,v_1),(v_2,v_3),(v_1,v_2)]$ qui implique la création d'un lien isolé au temps $t = 2$.

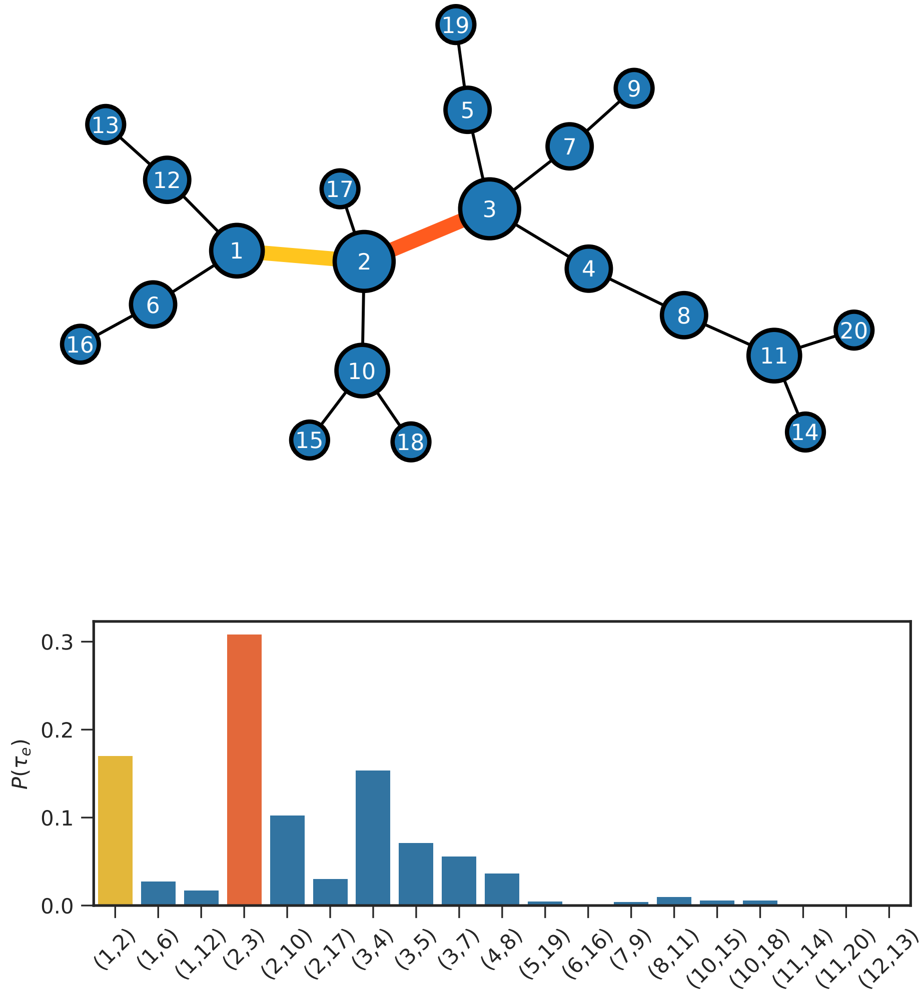


FIGURE 1.4 – (haut) Réseau artificiel de $N = 20$ noeuds, généré par le processus d'attachement préférentiel avec un exposant $\gamma = 0$. Les chiffres apparaissant sur les noeuds indiquent leur ordre d'arrivée. (bas) Estimation empirique de $P(\tau_e = 1)$, i.e., la probabilité que le lien e soit apparu le premier. La probabilité empirique est calculée à partir de $n = 20\,000$ ordonnements, générés par exploration aléatoire du réseau [voir Eq. (1.75)]. La méthode de rejet ne retient aucun échantillon en autant d'essais, et ne peut donc pas être montrée en comparaison.

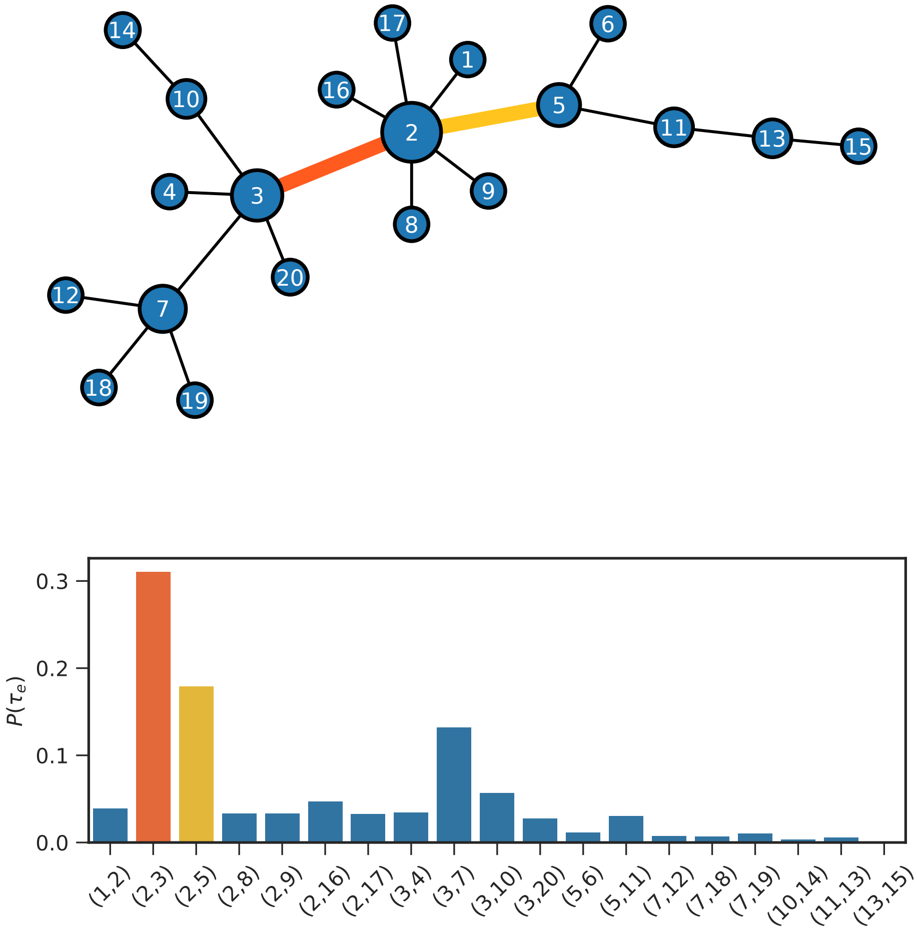


FIGURE 1.5 – (haut) Réseau artificiel de $N = 20$ noeuds, généré par le processus d'attachement préférentiel avec un exposant $\gamma = 1$. Les chiffres apparaissant sur les noeuds indiquent leur ordre d'arrivée. (bas) Estimation empirique de $P(\tau_e = 1)$, i.e., la probabilité que le lien e soit apparu le premier.
Le modèle assigne une faible probabilité $P(\tau_{(1,2)} = 1)$ au véritable premier lien. Il ne s'agit pas d'une erreur de l'échantillonneur, mais plutôt d'une variation aléatoire due au processus de croissance. En effet le lien $(1,2)$ ne joue pas le rôle structurel attendu. Le lien $(2,3)$ est plus central et mieux connecté aux noeuds de haut degré ; le modèle assigne donc une plus grande probabilité a posteriori $P(\tau_{(2,3)} = 1)$ à ce lien, avec raison. On note que ce genre de fluctuation est en fait systématique, puisqu'il y a une probabilité $1/2$ que le rôle structurel des deux premiers liens soit interchangé, par symétrie.

préférentiel assigne une probabilité $p \approx 0.702$ à l'événement

$$\text{«Le premier lien est parmi } \{(1,2),(2,3),(3,4),(3,5)\}.\text{»}$$

qu'on sait être vrai, puisque le lien $(1,2)$ est le premier. On obtient des résultats similaires dans le cas $\gamma = 1$ (Fig. 1.5).

1.3.4 Méthodes de Monte-Carlo par chaînes de Markov

Not all who wander are lost.
—J. R. R. Tolkien

On termine ce chapitre avec les méthodes de Monte-Carlo par chaînes de Markov (MCMC), probablement la classe de méthodes la plus populaire, en raison de sa polyvalence et de sa relative simplicité. Comme toujours, notre objectif est d'échantillonner une variable aléatoire X. Grossièrement, l'idée des méthodes MCMC est de (a) sélectionner un point de départ quelconque dans $\mathcal{X}$, le support de X, (b) de générer une chaîne d'éléments

$$x_0 \to x_1 \to x_2 \to \ldots \to x_T$$

en voyageant aléatoirement dans cet espace, et (c) construire l'échantillon en retenant périodiquement des éléments de la chaîne. Si les transitions sont aléatoires et que la chaîne satisfait quelques propriétés bien définies (irréductibilité, apériodicité), alors on peut traiter deux éléments de la chaîne séparés par un grand nombre d'étapes comme indépendants. Si la chaîne est à l'équilibre, alors ces éléments seront identiquement distribués, selon la distribution à l'équilibre $\boldsymbol{\pi}$ de la chaîne. Couplant ces deux observations, on voit que les méthodes MCMC permettent d'échantillonner des éléments i.d.d. de la distribution à l'équilibre du processus de marche aléatoire dans $\mathcal{X}$. L'ingéniosité des méthodes MCMC réside dans le fait qu'on peut choisir les probabilités de transition d'un point à l'autre de $\mathcal{X}$ de façon à ce que la distribution à l'équilibre de la marche aléatoire $\boldsymbol{\pi} = [\pi_1, ..., \pi_N]$ corresponde à la distribution à échantillonner. On présentera ici l'idée dans un espace $\mathcal{X}$ discret, mais l'approche MCMC fonctionne également pour des espaces continus [11].

La spécification d'une chaîne de Markov passe par la spécification d'une matrice de transition, reliant les points de l'espace $\mathcal{X}$. On note cette matrice $\boldsymbol{T}$; l'élément $T_{ji} \in [0,1]$ spécifie la probabilité de passer de l'état x_i vers l'état x_j. Notre but sera de choisir les éléments de $\boldsymbol{T}$ de façon à obtenir la distribution $\boldsymbol{\pi}$ désirée. Il y a plusieurs façons de faire ; on présente ici une approche standard, due à Metropolis et Hastings, qui est basée sur la condition de la balance détaillée :

$$T_{ji}\pi_i = T_{ij}\pi_j \qquad \forall\ (i,j)\ . \tag{1.76}$$

Dans une chaîne de Markov qui respecte la balance détaillée à l'équilibre, toutes les transitions ont la même probabilité que leur inverse. Il est facile de vérifier que si cette condition est respectée pour une distribution π quelconque, alors cette distribution est la distribution à l'équilibre : en sommant sur j et en utilisant la normalisation par colonne $\sum_j T_{ji} = 1$, conséquence de la définition des transitions, on obtient $\boldsymbol{\pi} = \boldsymbol{T}\boldsymbol{\pi}$.

L'idée de Metropolis et Hastings est de réécrire l'Eq. (1.76) en divisant les probabilités de transitions T_{ji} en deux parties : une distribution de proposition $q_{j\leftarrow i}$ (probabilité de proposer x_j dans l'état x_i) et une probabilité d'acceptation $\mathcal{A}_{j\leftarrow i}$ (probabilité d'accepter la transition proposée). On obtient

$$\frac{\pi_j}{\pi_i} = \frac{T_{ji}}{T_{ij}} = \frac{q_{j\leftarrow i}\mathcal{A}_{j\leftarrow i}}{q_{i\leftarrow j}\mathcal{A}_{i\leftarrow j}}. \tag{1.77}$$

Puisqu'on a débuté avec la balance détaillée, qui garantit que la distribution à l'équilibre est bien $\boldsymbol{\pi}$, on obtient la chaîne désirée en sélectionnant une paire de distributions $(q, \mathcal{A})$ qui respecte l'Eq. (1.77). Le choix standard est

$$\mathcal{A}_{j\leftarrow i} = \min\left\{1, \frac{\pi_j}{\pi_i}\frac{q_{i\leftarrow j}}{q_{j\leftarrow i}}\right\}\ , \tag{1.78}$$

puisque (a) la distribution d'acceptation est alors valide autant dans le cas continu que discret, (b) la normalisation de $\boldsymbol{\pi}$ n'est pas nécessaire car cette distribution apparaît comme un ratio de probabilité[36], et (c) l'analogie est facile avec le recuit simulé [124] (on le décrira sous peu).

Ce dernier avantage permet en fait d'obtenir les estimateurs MMSE et MAP à l'aide d'un même algorithme. En effet, tel que défini ci-haut, l'algorithme de *Metropolis-Hastings* échantillonne $\boldsymbol{\pi}$ et permet donc d'évaluer des estimateurs MMSE via le principe de Monte-Carlo. On peut cependant en faire un algorithme d'optimisation en pportant une petite modification à $\mathcal{A}_{i\leftarrow j}$: il suffit d'introduire un exposant β sur la distribution à l'équilibre :

$$\mathcal{A}^*_{j\leftarrow i}(\beta) = \min\left\{1, \frac{\pi_j^\beta}{\pi_i^\beta}\frac{q_{i\leftarrow j}}{q_{j\leftarrow i}}\right\}\ . \tag{1.79}$$

À chaque choix de $\beta > 0$ correspond une distribution $\pi'_i(\beta) \propto \pi_i^\beta$, concentrée sur les maxima de $\boldsymbol{\pi}$ lorsque $\beta \gg 1$. On peut donc trouver les maxima de la distribution d'origine en échantillonnant $\boldsymbol{\pi}'(\beta)$ dans la limite des grands β. Il faut cependant faire attention : on ne peut pas directement échantillonner avec $\beta \to \infty$, puisque l'algorithme est alors inutile ! En effet, l'algorithme cherche alors le maximum par échantillonnage uniforme des solutions, ce qui est très inefficace. Le principe du recuit simulé ; contourne le problème en combinant exploration et maximisation [155]. L'idée est d'alterner équilibration et optimisation, en augmentant tranquillement le paramètre β à partir d'une valeur initiale β_0 quelconque. En itérant ce procédé assez longtemps, on donne le temps aux distributions intermédiaires de se concentrer

[36] Une propriété qui est utile, e.g., quand on échantillonne un ERGM pour lequel $Z(\boldsymbol{\beta})$ est difficile à calculer.

sur les maxima, mais également d'explorer l'espace complet. Cette procédure mène ultimement à un meilleur résultat.[37]

Notons que la construction de Metropolis-Hastings ne spécifie pas la distribution $q_{i\leftarrow j}$; celle-ci sera dépendante du contexte, et déterminera largement l'efficacité de l'algorithme. Quelques considérations doivent rentrer en compte dans le choix d'une bonne distribution de proposition. Elle devra :

1. favoriser un mélange rapide (décorrélation rapide de la chaîne) ;
2. pouvoir explorer *tout* l'espace ;
3. être facilement calculable.

Un mélange rapide permet de retenir des échantillons plus souvent, ce qui accélère le calcul. Le deuxième point nous garantit une chaîne dite ergodique (qui peut explorer l'espace en entier, aléatoirement). Sans cette propriété, les échantillons seront forcément biaisés et la chaîne ne vaudra rien. Finalement, la dernière propriété garantit une implémentation numérique efficace.

Méthodes de Monte-Carlo par chaînes de Markov : Exemple

Pour clore le chapitre, on démontre l'utilité des méthodes MCMC en l'appliquant à la modélisation de réseau avec le SBM. Cette modélisation se déroule normalement comme suit. On commence avec un réseau G observé, qu'on pense être bien décrit par le modèle[38]. On peut par exemple se douter que les noeuds du réseau soient ségrégués en blocs qui trahissent leur fonction (e.g., dans le réseau des dauphins, on pourrait penser qu'il y a des groupes sociaux homogènes qui se fréquentent peu). Notre but est de trouver ces groupes et les densités associées, à partir de G uniquement.

Il s'agit d'un problème d'estimation de paramètres avec le SBM : la structure cachée du réseau est contenue dans $\hat{\boldsymbol{\sigma}}$ et la caractérisation de ces blocs est donnée par $\hat{\boldsymbol{P}}$[39]. On veut donc calculer les estimateurs de la distribution conjointe

$$P(\boldsymbol{\sigma},\boldsymbol{P}|G)=\frac{P(G|\boldsymbol{\sigma},\boldsymbol{P})P(\boldsymbol{P},\boldsymbol{\sigma})}{P(G)}\ ,$$

un problème assez difficile parce les deux ensembles de paramètres interviennent conjointement dans le calcul de la distribution a posteriori. Deux stratégies permettent de simplifier le problème :

[37] Le terme recuit simulé (*simulated annealing*) provient d'une analogie avec la physique de l'état solide, où un refroidissement lent d'un matériel lui permet de trouver un minimum d'énergie interne très faible. L'augmentation de β simule le refroidissement, et la position de la chaîne dans $\boldsymbol{X}$ donne son «énergie» $[\pi(x_i)]^{-1}$. Le refroidissement force la chaîne vers ses états de basse énergie.

[38] C'est en fait toujours le cas. Le SBM peut être vu comme une approximation universelle [180].

[39] Pour simplifier la discussion, on suppose que q est un paramètre fixe. Voir par exemple [177, 192] pour des méthodes permettant d'estimer ce paramètre.

1. l'estimation alternée des paramètres (échantillonnage de Gibbs et inférence via l'algorithme d'espérance-maximisation [153]) ;
2. l'inférence via le modèle intégré $\int P(\boldsymbol{\sigma},\boldsymbol{P}|G)d\boldsymbol{P}$.

On a vu dans la section 1.2.2 que la deuxième option est approximativement équivalente à estimer les paramètres du SBM microcanonique. On préférera ici la première approche, qui permet d'apprendre les deux ensembles de paramètres et de mieux illustrer notre propos. Voir les références [190, 193] pour une discussion de la version microcanonique.

Le but de l'estimation alternée est de converger vers les points les plus importants de la distribution a posteriori, en se déplaçant uniquement dans des sous-espaces de l'espace des paramètres. Plus explicitement, en commençant par exemple par fixer une configuration initiale quelconque $\hat{\boldsymbol{\sigma}}_0$ (normalement au hasard), on trouve $\hat{\boldsymbol{P}}_0$, pour cette configuration. On utilise ensuite $\hat{\boldsymbol{P}}_0$ pour estimer $\hat{\boldsymbol{\sigma}}_1$, et ainsi de suite. On peut montrer que cette méthode est justifiée dans la plupart des situations [11, 118]. Nous avons déjà calculé les estimateurs de $\boldsymbol{P}$ pour une assignation $\boldsymbol{\sigma}$ fixe, voir Eq. (1.58). On se concentre donc sur le problème complémentaire, soit celui d'évaluer

$$P(\boldsymbol{\sigma}|G,\hat{\boldsymbol{P}}) = \frac{P(G|\boldsymbol{\sigma},\hat{\boldsymbol{P}})P(\boldsymbol{\sigma})}{P(G)} \,. \tag{1.80}$$

C'est ici que l'algorithme MCMC entre en jeu : nous allons l'utiliser pour calculer $P(\boldsymbol{\sigma}|G,\hat{\boldsymbol{P}})$ numériquement.

Le support de $\boldsymbol{\sigma}$ est l'ensemble des assignations à q blocs non vides, un espace discret de $O(q^N)$ éléments. La distribution de proposition $q_{j\leftarrow i}$ doit permettre d'explorer cet espace. On opte pour une distribution qui possède cette propriété de façon évidente : à chaque pas de la chaîne, on propose de modifier l'assignation d'un seul noeud, choisi uniformément parmi tous les noeuds. Puisqu'il y a N noeuds et q blocs, on a

$$q_{j\leftarrow i} = q_{i\leftarrow j} = \frac{1}{Nq} \,, \tag{1.81}$$

de sorte que la probabilité d'acceptation est donnée par

$$\mathcal{A}_{j\leftarrow i} = \min\left\{1, \frac{P(\boldsymbol{\sigma}_j|G,\hat{\boldsymbol{P}})}{P(\boldsymbol{\sigma}_i|G,\hat{\boldsymbol{P}})}\right\} \,.$$

On choisit une loi a priori uniforme sur $\boldsymbol{\sigma}$, et on dénote par v l'indice du noeud sélectionné à l'étape de proposition. La probabilité d'acceptation est alors

$$\begin{aligned}\mathcal{A}_{j\leftarrow i} &= \min\left\{1, \frac{\prod\limits_k (1-p_{\sigma_i(v)\sigma_i(k)})^{1-a_{vk}}(p_{\sigma_i(v)\sigma_i(k)})^{a_{vk}} \prod\limits_{k<l:k\neq v,l\neq v}(1-\hat{p}_{\sigma_i(k)\sigma_i(l)})^{1-a_{kl}}(\hat{p}_{\sigma_i(k)\sigma_i(l)})^{a_{kl}}}{\prod\limits_k (1-p_{\sigma_j(v)\sigma_j(k)})^{1-a_{vk}}(p_{\sigma_j(v)\sigma_j(k)})^{a_{vk}} \prod\limits_{k<l:k\neq v,l\neq v}(1-\hat{p}_{\sigma_j(k)\sigma_j(l)})^{1-a_{kl}}(\hat{p}_{\sigma_j(k)\sigma_j(l)})^{a_{kl}}}\right\} \\ &= \min\left\{1, \prod_k \left(\frac{1-\hat{p}_{\sigma_i(v)\sigma_i(k)}}{1-\hat{p}_{\sigma_j(v)\sigma_j(k)}}\right)^{1-a_{vk}} \left(\frac{\hat{p}_{\sigma_i(v)\sigma_i(k)}}{\hat{p}_{\sigma_j(v)\sigma_j(k)}}\right)^{a_{vk}}\right\}\end{aligned} \tag{1.82}$$

où on a factorisé les normalisations $P(G)$ et où on a utilisé l'indépendance des liens pour retirer tous les termes n'impliquant pas directement le noeud v [40].

On applique l'algorithme complet (estimation en alternance) au réseau de dauphins, avec $q = 2$ blocs. Par souci de simplicité on utilise une version homogène du SBM, définie par des probabilités

$$p_{rs} = \begin{cases} p_{\text{in}} & \text{si } r = s, \\ p_{\text{out}} & \text{autrement.} \end{cases} \tag{1.83}$$

Le modèle homogène est plus simple et nous permet de visualiser l'évolution des paramètres via le ratio $\hat{p}_{\text{in}}/\hat{p}_{\text{out}}$. Puisqu'aucune information a priori n'est disponible, on initialise l'algorithme complet à $\hat{p}_{\text{in}}/\hat{p}_{\text{out}} = 1$ (un choix arbitraire), puis on cherche une première partition à l'aide de l'estimateur $\hat{\boldsymbol{\sigma}}^{\text{MMO}}$.

On rappelle que pour calculer cet estimateur (c.f. Sec. 1.2.2), on doit construire les distributions marginales $P(\sigma(i)|G, \hat{p}_{\text{in}}, \hat{p}_{\text{out}})$ pour chaque noeud $i = 1, ..., N$. L'estimateur MMO est obtenu en assignant chaque noeud à son groupe le plus probable. Comme on l'a annoncé, on calcule la distribution $P(\sigma(i)|G, \hat{p}_{\text{in}}, \hat{p}_{\text{out}})$ à l'aide de l'échantillonneur MCMC. Plus précisément, on initialise l'échantillonneur avec une partition aléatoire et on ignore les 5 000 premiers échantillons (puisqu'il est peu probable que la chaîne soit dans son état stationnaire dès le début). On retient ensuite une partition à chaque 100 itérations de la chaîne, $n = 5\,000$ fois, ce qui nous permet d'estimer la distribution a posteriori via le principe de Monte-Carlo :

$$\hat{P}(\sigma(i) = r|G, \hat{p}_{\text{in}}, \hat{p}_{\text{out}}) = \frac{1}{n}\sum_{k=1}^{n} \mathbb{I}[\sigma_k(i) = r]\,. \tag{1.84}$$

On prend ensuite $\hat{\sigma(i)}^{\text{MMO}} = \text{argmax}_r \hat{P}(\sigma(i) = r|G, \hat{p}_{\text{in}}, \hat{p}_{\text{out}})$ pour chaque noeud, ce qui complète le calcul de la partition avec des densités fixes ; on vient de terminer une itération de l'algorithme d'estimation alternée.

Pour poursuivre, on doit maintenant estimer les densités p_{in} et p_{out} afin de remplacer les paramètres initiaux (arbitraires). On a donc recours aux estimateurs MMSE de l'Eq. (1.58), avec la partition $\hat{\boldsymbol{\sigma}}^{\text{MMO}}$. En l'absence d'information a priori, on utilise des paramètres dits non informatifs, i.e., $\alpha_{rs} = \beta_{rs} = 1/2\ \forall\ (r,s)$ [209]. Puisqu'on utilise le modèle homogène, on obtient des paramètres simplifiés en calculant les moyennes $\hat{p}_{\text{in}} = (\hat{p}_{11} + \hat{p}_{22})/2$ et $\hat{p}_{\text{out}} = \hat{p}_{12}$. On est alors prêt à utiliser l'échantillonneur MCMC à nouveau.

Les résultats de cette procédure sont montrés à la figure 1.6. On trouve la même séparation que celle obtenue à l'aide du modèle microcanonique (voir Fig. 1.1), ce qui indique que le modèle homogène est probablement un bon choix pour décrire ce réseau ; il est plus simple, mais tout aussi explicatif que le SBM complet. Le panneau de droite de la figure indique

[40] En pratique, cette forme n'est pas utilisée car elle est trop coûteuse numériquement. On préfère une version condensée où les termes sont regroupés par blocs. L'équation explicite est donnée dans la section 3.9.5 du chapitre 3.

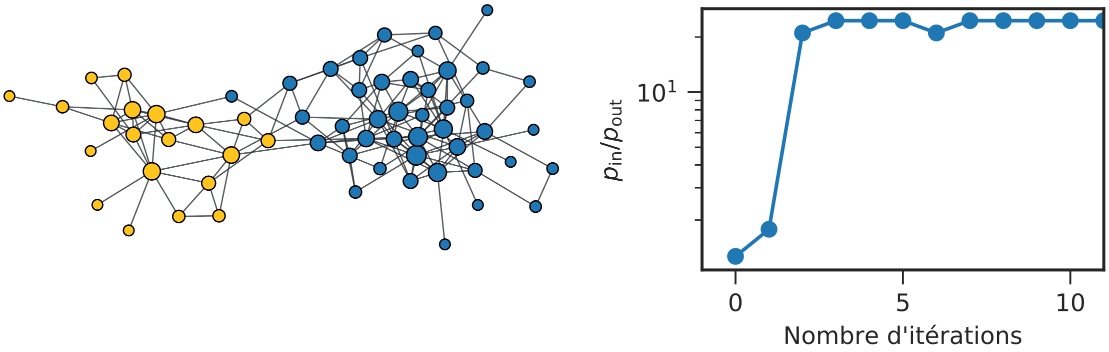


FIGURE 1.6 – Estimation de paramètres pour le SBM canonique. (gauche) Partition obtenue, identique à celle trouvée par SBM microcanonique, illustrée à la Fig. 1.1. (droite) Évolution des paramètres du modèle en fonction du nombre d'itérations. On utilise les conditions initiales $p_{\mathrm{in}} = p_{\mathrm{out}} = 0.5$. L'algorithme converge rapidement vers $p_{\mathrm{in}} = 0.17, p_{\mathrm{out}} = 0.007$.

que l'algorithme d'estimation alternée converge rapidement vers la valeur finale des paramètres. On peut donc faire confiance aux résultats. Soulignons que ces résultats sont obtenus en considérant 5×10^4 configurations σ, une fraction infinitésimale des $q^N \approx 4.6 \times 10^{18}$ configurations σ possibles : un tour de force !

Première partie

Structure mésoscopique

Chapitre 2

Structure mésoscopique I : De l'universalité du modèle stochastique par blocs

Article original :

"Universality of the stochastic block model"

Jean-Gabriel Young [1, 2], Guillaume St-Onge [1, 2], Patrick Desrosiers [1, 2 ,3] et Louis J. Dubé [1, 2]

[1] Département de Physique, de Génie Physique, et d'Optique
Université Laval, Québec (QC), G1V 0A6, Canada
[2] Centre interdisciplinaire de modélisation mathématique de l'Université Laval
Québec (QC), G1V 0A6, Canada
[3] Centre de recherche de CERVO, Québec (QC), G1J 2G3, Canada

Référence : *Phys. Rev. E* (sous presse) et arXiv:1806.04214 [251].

‡ Ces sections sont reproduites directement de l'article orignal. Le contenu n'en a été modifié que pour se conformer au format exigé par la Faculté des études supérieures et postdoctorales de l'Université Laval.

2.1 Avant-propos

On se rappellera que dans le premier chapitre, on a utilisé le modèle stochastique par blocs (SBM) pour identifier des groupes cohésifs de noeuds dans un réseau social (c.f. Fig. 1.1). Ce faisant, on a identifié la *structure mésoscopique* de ce réseau, i.e., son organisation à un niveau de résolution quelque part entre le niveau local (voisinage immédiate des noeuds) et le niveau global (réseau en entier). On a ainsi trouvé des régularités, nous permettant de résumer sa structure.

Ce problème—l'identification de régularités—est parmi les mieux étudiés de la science des réseaux. Il existe donc maintenant une foule d'algorithmes, souvent *ad hoc*, visant à découvrir différents types de régularités [78] : blocs densément connectés, séparations en un coeur et une périphérie, etc. Or, depuis quelques années, on comprend que le SBM est plus polyvalent que la plupart de ces méthodes [1, 171, 180, 190, 197], puisqu'il n'y a que très peu de restrictions *a priori* sur les régularités pouvant être extraites par le modèle. Dans le présent chapitre, on montre que cette polyvalence s'explique par le fait que le SBM peut être vu comme une méthode de détection *universelle*.

2.2 Résumé

L'extraction de régularités mésoscopiques (ERM) désigne tout problème consistant à identifier une partition des noeuds d'un réseau complexe qui maximise une fonction objective quelconque. Plusieurs problèmes d'inférence réseau font partie de cette catégorie, tels que : détection de communautés, séparation en coeur–périphérie, coloration imparfaite de graphes, etc. Dans le présent chapitre, on montre que la plupart des algorithmes solutionnant des problèmes d'ERM peuvent être vus comme maximisant la vraisemblance de cas spéciaux du modèle stochastique par blocs (SBM), ou d'une de ses généralisations directes. La découverte de ces relations d'équivalences montrent donc que le SBM est un modèle quasi-universel pour l'ERM.

2.3 Abstract

Mesoscopic pattern extraction (MPE) is the problem of finding a partition of the nodes of a complex network that maximizes some objective function. Many well-known network inference problems fall in this category, including, for instance, community detection, core-periphery identification, and imperfect graph coloring. In this paper, we show that the most popular algorithms designed to solve MPE problems can in fact be understood as special cases of the maximum likelihood formulation of the stochastic block model (SBM), or one of its direct generalizations. These equivalence relations show that the SBM is nearly universal with respect to MPE problems.

2.4 Introduction

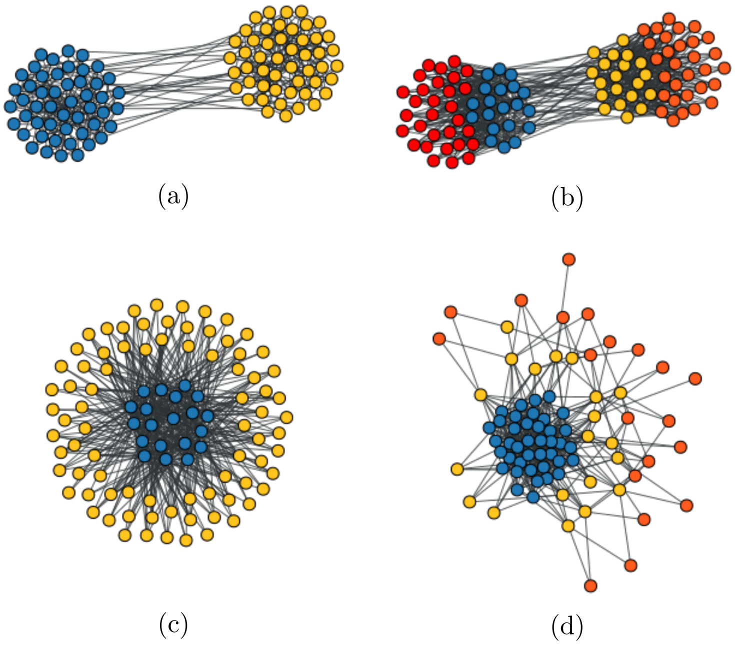
Whether it is called community detection, graphical inference, spectral embedding, unsupervised learning, bisection or graph coloring, the idea of summarizing the structure of a complex system by grouping its elements in blocks is a popular one, discovered time and time again in different areas of science [158]. As such, there are now a plethora of algorithms and techniques—developed essentially in parallel—that provide good solutions to this ubiquitous problem [78]. In the past few years, a great deal of work has been done toward unifying and contrasting these approaches, building bridges across cultural divides [158, 216]. This has been fruitful work thus far, for—sometimes surprising—equivalences between drastically different methods have turned up in the process, e.g., between modularity and the maximum likelihood formulation of the degree-corrected stochastic block models (SBM) [174, 214, 234, 254], various spectral methods [173], normalized-cut [120], random-walks [151], and non-negative matrix factorization [39]. These results invite the question : Is there a deeper reason for the correspondences, or are they simply mathematical coincidences ?

The purpose of this paper is to show that equivalences arise because most of these *mesoscopic pattern extraction* (MPE) methods are actually the maximum likelihood formulation of the SBM in disguise (and a generalization of its degree–corrected version [13, 116]). By MPE problems, we mean any problem where one is asked to find a partition of the network that maximizes some implicit or explicit score, encoded via an objective function.

Our results rest on the concepts of *equivalence* and *specialization* of the objective functions : Two objective functions are equivalent when they order any pair of partitions the same way (i.e., they implement the same notion of optimality), and specialization refers to the idea of limiting the expressiveness of an objective function by fixing some of its parameters (see Sec. 2.5). With these two operations, we delineate a hierarchy that crystallizes the idea of the SBM as a general MPE tool : Through specialization of its likelihood, it can be tailored to find patterns such as assortative and disassortative communities [174], bipartite structures [133], or core-periphery splits [28] (in Sec. 2.6). Importantly, we show that these specialized likelihoods are *exactly* equivalent to the objective functions implemented by MPE algorithms such as modularity maximization, balanced cut, core-periphery search, etc. Our framework therefore offers principled methods to determine any arbitrary parameters that might arise in otherwise ad hoc modularities [174], but also suggests statistical techniques to carry out principled inference, in the spirit of Refs. [229, 254] (see Sec. 2.7).

2.5 Mesoscopic structures and optimization

The mesoscopic pattern extraction (MPE) problem is usually stated as follows. We are given an extremely large complex network, generated by some random hidden process. Its overall organization is impossible to grasp, because its structure is much too detailed. Our goal



(a)

(b)

(c)

(d)

FIGURE 2.1 – Four examples of mesoscopic pattern extraction problems on artificial networks. (a) Community detection [76, 176], (b) community detection with further structure, (c) the identification of a simple core-periphery split [28, 210], (d) identification of a core with a structured periphery. The targeted patterns are identified with colors.

with MPE is to reduce this complexity, by subsuming nodes in larger coherent units, using the structure of the network as our only input (and possibly additional metadata [108, 175]). Sometimes, the hope is to reveal functional components and hints about assembly mechanisms, while at other times it is only a matter of making the dataset more manageable, or interpolating from what is known [12, 25, 41, 76, 78, 203, 216, 220]. There is, however, a common theme : MPE algorithms take a complex network as their input, and produce as output a *partition* $\mathcal{B} = \{B_1, ..., B_q\}$ of the n nodes in q blocks $B_1, ..., B_q$, assigning precisely one block to each node. Despite these commonalities, the definition of what is a suitable partition will of course depend on the MPE problem at hand ; there is thus a wealth of MPE *algorithms*, reflecting the wealth of MPE *problems* (see Fig. 2.1).

To establish parallels between algorithms of diverse natures, we must first clearly answer : What is the essence of an MPE algorithm ? And what do we mean, when we say that two algorithms are *equivalent* ? The answers to these questions are not trivial, and crucial to the interpretation of the results of Secs. 2.6.1–2.6.2. Our goal with the next four subsections is therefore to clarify these issues.

2.5.1 Anatomy of a black box

There are essentially two possible ways to formulate our answers, depending on how we think of MPE algorithms.

First, we may take the empirical point of view and declare that the essence of algorithms is their *action*, independent of their inner workings. According to this point of view, equivalence is functional and context dependent : If two algorithms give the same result on a series of networks $G_1, G_2, ...G_k$, then the algorithms are equivalent with respect to these k networks. This allows us to treat algorithms as black boxes : Network in, partition out. It is certainly an appealing approach, because it may be used to compare algorithms of widely different natures—say a genetic algorithm with an evolved objective function and a label propagation method. Functional equivalence, however, has the drawback that it depends on the context, which makes it hard to draw definitive conclusions about algorithms. Furthermore, it may identify somewhat artificial parallels, because it is insensitive to the *origin* of the equivalences.

A second point of view is centered on the *definitions* of MPE algorithms rather than their action therefore appears necessary. Due to the diversity of existing MPE algorithms, this point of view will only be useful if we are able to first express MPE algorithms in some canonical form that can be readily analyzed. One possibility is a two–part model expressed as the coupling of (i) an objective function that induces a total ordering of the partitions, and (ii) a maximizer that can find a—potentially local—optimum of the objective function (see Fig. 2.2). This two–part model captures the two important mechanisms that any MPE algorithm must possess. On the one hand, the objective function captures the notion of quality of the partition and, consequently, tells the algorithm when to stop, and what partition to prefer whenever it has a choice. On the other hand, the maximizer provides a mean of moving in the solution space, and of pinpointing the best partitions, as per the above criterion. These mechanisms might be interwoven or hidden—we will touch on the subject shortly—, but the separation holds quite generally.

With the two–part model in place, equivalence takes on a crisp and clear meaning. Two algorithms are either partially equivalent—same objective **or** same maximizer—or completely equivalent—same objective **and** same maximizer. In the present paper, we will focus on partial equivalence, essentially ignoring the maximizers. This choice is motivated by the observations that (a) maximizers are, by necessity [1], only efficient heuristics designed to find "good enough" optima in the rugged landscape of partitions [158, 189, 227] (b) the no free lunch theorem implies that different objective functions and different inputs are associated with different optimal maximizers [189]. Hereafter, by *equivalence*, we will therefore refer to the equivalence of the objective functions used.

2.5.2 A glance under the hood

In the simplest—and quite common—case, the separation in two parts is explicit. For example, modularity–based methods famously attempt to maximize the modularity function over the

[1] MPE problems are quite generally in NP-HARD [31, 50].

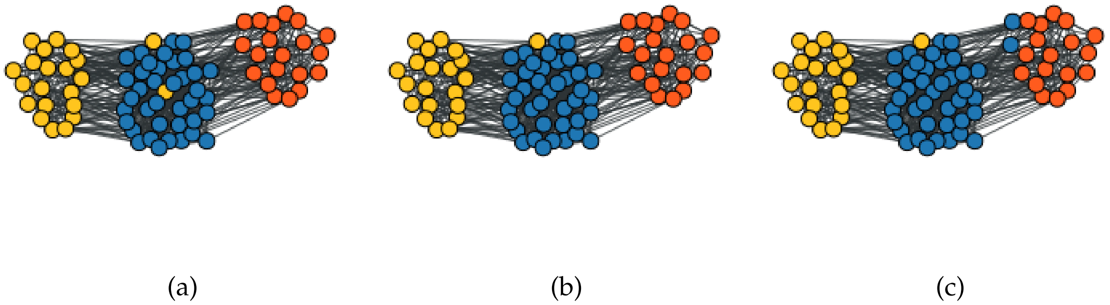


FIGURE 2.2 – Two–part algorithms in practice. We apply different combinations of maximizers and objective functions to a simple network with a clear block structure, generated using the SBM. Different MPE algorithms reveal different mesoscopic structures, but changing the objective function has the largest impact. (a, b) Same objective functions (modularity, see Sec. 2.6.2) and different maximizers [(a) spectral, (b) greedy]. (b, c) Same maximizer (greedy) with two different objective functions [(b) modularity, (c) core-periphery see Sec. 2.6.1]. The greedy maximizer is adapted from the Kernighan–Lin algorithm [123], in the spirit of Ref. [116], and the spectral maximizer relies on the embedding of the $q-1$ leading eigenvectors of the modularity matrix [169, 176] in $\mathbb{R}^{q-1}$, followed by a clustering step, here implemented using a Gaussian mixture [239].

set of all partitions of a network [76]. If there are many modularity optimization *algorithms*, it is because there are many different mechanisms that can propose and refine partitions to find the optima of the modularity, e.g., the iterative spectral method of Ref. [169], the fast unfolding method of Ref. [25], or the message-passing algorithm of Ref. [254]. The two–part algorithmic model is an exact description of these methods because they are framed in the language of objective functions.

Importantly, the two–part algorithmic model also holds in many cases where the emphasis is shifted away from an explicit objective function and maximizer dichotomy. Consider as an example the classical label propagation algorithm of Ref. [206]. This algorithm moves through the partition space by first assigning temporary labels (blocks) to nodes, and then repeatedly updating the labels with a majority rule (a node takes the label worn by the majority of its neighbors). Optimality is thus not defined for arbitrary pairs of partitions ; it is instead expressed as a dynamical, initial condition dependent concept. But a description in two parts can still be given, provided that we do some translation work : The label propagation mechanism can be thought of as a maximizer, which naturally leads to *partition flow* as a notion of optimality. A partition $\mathcal{B}_1$ is better than $\mathcal{B}_2$ if the algorithm goes from $\mathcal{B}_2$ to $\mathcal{B}_1$ when it updates labels based on the majority rule. With this definition, the best partitions are those that are stable against majority updates, and they are found via the propagation of labels. One can construct an objective function with these orderings, and therefore a two–part algorithm indistinguishable from the original [17, 231, 235].

2.5.3 General graphical objective function

Having established that a separation of algorithms into an objective function and a maximizing mechanism is often possible, let us turn to the functions themselves.

The outcome of pairwise interactions determines the structure of a complex network. A *general* objective function devised to uncover the mesoscopic patterns of a network therefore ought to include all these interactions in its calculation, at the very least. If it does no more than that, then the function can be called *graphical*, in the sense that no high-order terms are considered (i.e., there are no direct dependency on triplet of nodes, etc.). From this point onward, we will focus on graphical objective functions alone; the remainder of this paper is a testament to the generality of such a "limited" approach.

The definition of graphical objective function begins with the definition of its basic elements : Scores associated with each pair of nodes. For the sake of generality, we will define these scores as real-valued functions, with essential dependencies on the partition $\mathcal{B} = \{B_1, \ldots, B_q\}$ under consideration, on the structure of the network as encoded by the $n \times n$ adjacency matrix $\boldsymbol{A}$, and on an additional $n \times n$ side-information matrix $\boldsymbol{\lambda}$ that contains any pairwise information not directly captured by $\boldsymbol{A}$. Let us therefore write the score associated with the pair of nodes $(i,j) \in [n] \times [n]$ (we use the integers $[k] = \{1,...,k\}$ to denote the nodes) as

$$f(a_{ij}, \lambda_{ij}, \sigma_i, \sigma_j), \tag{2.1}$$

where $\sigma_i \in [q]$ is the index of the block of node i, i.e., $\sigma_i = r$ if and only if $i \in B_r$. We then express the aggregate of these local scores as

$$H\big(\boldsymbol{A}, \boldsymbol{\lambda}, \sigma; f\big) = \sum_{i,j: 1 \leq i \leq j \leq n} f(a_{ij}, \lambda_{ij}, \sigma_i, \sigma_j)\,, \tag{2.2}$$

yielding a global objective function based on pairwise scores. This form highlights the close parallel that exists between graphical objective functions and Edward–Anderson Hamiltonians [227], explicitly harnessed in a number of specific cases in Refs. [56, 124, 207, 208, 211], for example. The choice of a sum is, otherwise, for mere convenience; a product aggregate could have been equivalently implemented by taking $f \mapsto \log f$ and $H \mapsto e^H$.

2.5.4 Equivalence and hierarchy under specialization

The last piece of the theoretical framework is a clear notion of connections among functions. We use two concepts to establish these connections : Equivalence and specialization.

Equivalence

We say that two objective functions are equivalent if they induce the same total ordering of partitions, regardless of their inputs. This definition captures the correct notion of equivalence, because it is clear that two equivalent objective functions—by this standard—will yield two MPE indistinguishable algorithms when they are paired with the same maximizer. As it stands, however, this notion of equivalence is not easy to handle mathematically. We therefore resort to a second, stronger, criterion that leads to a more direct comparison procedure. It is obvious that if

$$H(\boldsymbol{A},\boldsymbol{\lambda},\boldsymbol{\sigma};f) < H(\boldsymbol{A},\boldsymbol{\lambda},\boldsymbol{\sigma}';f) \implies g \circ H(\boldsymbol{A},\boldsymbol{\lambda},\boldsymbol{\sigma};f) < g \circ H(\boldsymbol{A},\boldsymbol{\lambda},\boldsymbol{\sigma}';f)\ , \tag{2.3}$$

for some strictly increasing function g, then H and $g \circ H$ are equivalent according to the first definition. While this second version is more restrictive, it reduces the comparison of objective functions to the identification of the transformation g—an often straightforward process.

As we will see in Sec. 2.6, in practice, an even stronger criterion that limits g to a particular subset of all *linear* transformation will often suffice to establish many equivalence relations. Namely, whenever a pairwise score functions $f(a_{ij},\lambda_{ij},\sigma_i,\sigma_j)$ can be split as

$$f(a_{ij},\lambda_{ij},\sigma_i,\sigma_j) = f_1(a_{ij},\lambda_{ij},\sigma_i,\sigma_j) + f_2(a_{ij},\lambda_{ij}), \tag{2.4}$$

where f_2 does not depend on the partition, we will be able to rewrite the global objective function as

$$\begin{aligned} H(\boldsymbol{A},\boldsymbol{\lambda},\boldsymbol{\sigma};f) &= \sum_{i\leq j} f_1(a_{ij},\lambda_{ij},\sigma_i,\sigma_j) + \sum_{i\leq j} f_2(a_{ij},\lambda_{ij}) \\ &\sim \sum_{i\leq j} f_1(a_{ij},\lambda_{ij},\sigma_i,\sigma_j) \\ &= H'(\boldsymbol{A},\boldsymbol{\lambda},\boldsymbol{\sigma};f) \end{aligned} \tag{2.5}$$

where "$\sim$" denotes equivalence, and "$i \leq j$" is a shorthand for the more precise statement "$i,j : 1 \leq i \leq j \leq n$." The equivalence holds because the additive terms are independent from $\boldsymbol{\sigma}$ and therefore do not affect the ordering. Thus, equivalence will often follow from a simple linear transformation of the form $g \circ H = H - \sum_{i\leq j} f_2(a_{ij},\lambda_{ij})$.

Specialization

With specialization, we aim to capture the idea that an objective function can be less expressive than its *parent* function, i.e., that it is possible to fix some parameters of a function (the parent) to obtain a "simpler" version of the function[2]. It is more straightforward to define specialization at the level of pairwise score, and so we will say informally that a pairwise score function f_S is a specialization of f if f_S is constructed by fixing some of the free parameters of f, in a way that alters the ranking of partitions, for some inputs. Furthermore, we

[2] Formally, f is in fact a placeholder for a function space with some parametrization set $\boldsymbol{\pi}$; f only represent a unique function upon choosing some $p \in \boldsymbol{\pi}$. Specializing f corresponds to defining a function space f_S associated with a parameter set $\boldsymbol{\pi}_S \subseteq \boldsymbol{\pi}$.

will say that the objective function $H'(\boldsymbol{A},\boldsymbol{\lambda},\boldsymbol{\sigma};f_S)$ is a specialization of the objective function $H(\boldsymbol{A},\boldsymbol{\lambda},\boldsymbol{\sigma};f)$ when f_S is a specialization of f.

In the context of MPE, if f_S is derived from f and there exists at least one pair of nodes (i,j) such that

$$\begin{gathered} f_S(a_{ij},\lambda_{ij},\sigma_i,\sigma_j) = f_S(a_{ij},\lambda_{ij},\sigma_i',\sigma_j') \\ \text{and} \\ f(a_{ij},\lambda_{ij},\sigma_i,\sigma_j) \neq f(a_{ij},\lambda_{ij},\sigma_i',\sigma_j')\;, \end{gathered} \tag{2.6}$$

where $\boldsymbol{\sigma} \neq \boldsymbol{\sigma}'$, then f_S is a specialization of f (and similarly for the resulting H' and H).

Specialization is, in a sense, a one-way operation, because it involves reducing the complexity of a function. In Eq. (2.6), f could act as f_S but not the other way around, because f_S is derived from f by specialization. Thus, specialization induces a hierarchy, with the most general functions at the top, and the most specialized ones at the bottom. This is the hierarchy that we propose to delineate in the next sections.

2.6 Objective function hierarchy under specialization

Recall that our claim is essentially the following : The objective functions of many mesoscopic pattern extraction algorithms are, in fact, special cases of the maximum likelihood formulation of the SBM. Sections 2.6.1 and 2.6.2 are devoted to showing how this comes about. We begin with the methods that *do not* account for any side information λ_{ij}, in Sec. 2.6.1. We show that they can be understood as specialization of the maximum likelihood formulation of the classical SBM [105]. We then move on to general MPE methods, in Sec. 2.6.2, by adding a side information dependency to the score functions. Again, we show that these methods can be seen as specializations of a generalized SBM, close in spirit to the degree–corrected SBM of Ref. [116]. This part of the hierarchy sits above the methods of Sec. 2.6.1, since the generalized SBM contains the classical SBM as a special case. We summarize the relations between the various methods in Fig. 2.3, and show in Fig. 2.4 that they can be used to extract various patterns from a same real network.

2.6.1 Partial hierarchy (no side information)

Stochastic block model

Our starting point is the stochastic block model (SBM). It is not an MPE algorithm *per se*, but rather a random network model, amenable to statistical inference. It prescribes a likelihood for the network G, parametrized by a latent partition $\mathcal{B}$ of its nodes. The SBM becomes a MPE algorithm once this likelihood is used to infer the hidden partition $\mathcal{B}$ of G. Although there are many ways of harnessing the likelihood to extract the mesoscopic patterns encoded by $\mathcal{B}$, we will only focus on *likelihood maximization*, because it directly fits within the two–part model of MPE algorithms defined in Sec. 2.5.1; the likelihood is the objective function and the maximizer does not matter.

Given a network and a partition of the nodes in blocks associated with the vector $\boldsymbol{\sigma}$, the classical SBM[3] prescribes that the number of edges between nodes (i,j) should be drawn from a Poisson distribution of mean $\omega_{\sigma_i\sigma_j}$. All edges are assumed to be independent, such that the likelihood of the complete graph is given by

$$\mathbb{P}(G|\mathcal{B},\boldsymbol{\omega}) = \prod_{i\leq j} \frac{(\omega_{\sigma_i\sigma_j})^{a_{ij}}}{a_{ij}!} e^{-\omega_{\sigma_i\sigma_j}} \ . \tag{2.7}$$

It is parametrized by the $q \times q$ matrix $\boldsymbol{\omega}$ and the partition $\mathcal{B}$ (or equivalently by the block assignments $\boldsymbol{\sigma}$). The standard inference procedure calls for the estimation of both, $\boldsymbol{\omega}$ and $\mathcal{B}$, usually through alternated learning of the two sets of parameters (via the expectation–maximization algorithm [56]). However, we will focus on the estimation of $\mathcal{B}$ alone, treating the parameters $\boldsymbol{\omega}$ as "control buttons." The freedom to *impose* parameters $\boldsymbol{\omega}$ on the network will ultimately allow us to draw relations with other MPE algorithms.

To extract $\mathcal{B}^*(G)$—the "true" partition of the nodes—from the network, we maximize the likelihood of the SBM with respect to the partition (see also Sec. 2.7). Since the logarithm is a strictly increasing function of its argument, we may equivalently maximize the log-likelihood

$$\log \mathbb{P}(G|\mathcal{B},\boldsymbol{\omega}) = \sum_{i\leq j} \Big[a_{ij} \log \omega_{\sigma_i\sigma_j} - \omega_{\sigma_i\sigma_j} - \log a_{ij}! \Big] \ . \tag{2.8}$$

This is a first (trivial) example of the concept of equivalence of Sec. 2.5.4. It becomes evident upon inspection of Eq. (2.8) that the log-likelihood is, in fact, a graphical objective function of the general form appearing in Eq. (2.2), associated with the pairwise score function

$$f_{\mathrm{SBM}}(a_{ij},\sigma_i,\sigma_j) \sim a_{ij} \log \omega_{\sigma_i\sigma_j} - \omega_{\sigma_i\sigma_j} \ . \tag{2.9}$$

Thus, any objective function that can be written as a special case of Eq. (2.9) will be a specialization of the maximum likelihood formulation of the SBM.

General modular graph model

One such (explicit) specialization is the general modular graph model (GMGM) [121, 247]. Like its general counterpart, the GMGM is a generative model for networks that supposes a latent partition of the nodes in blocks. The crucial difference is that the connection matrices $\boldsymbol{\omega}$ of the GMGM are much more structured than that of the SBM.

[3] We base our derivation on the Poisson SBM, nearly equivalent to the somewhat more standard Bernoulli SBM. Our choice is justified by the fact that the full hierarchy of objective functions follows more naturally from the Poisson SBM.

Pairs of blocks are assigned one of two types, say a and b, and this information is encoded in a $q \times q$ binary (and symmetric) matrix $\boldsymbol{X}$. If a pair of blocks (B_r, B_s) is of type a, then we set $x_{rs} = 1$. Contrariwise, we set $x_{rs} = 0$ if the pair (B_r, B_s) is of type b. Pairs of blocks of type a are then *all* associated with a connectivity $\omega_{rs} = \omega_a$, while pairs of type b are associated with a connectivity $\omega_{rs} = \omega_b$, where we take $\omega_b < \omega_a$ without loss of generality [4]. Every connection matrix of the GMGM can therefore be written as

$$\boldsymbol{\omega} = \omega_b \mathbf{1}\mathbf{1}^{\intercal} + (\omega_a - \omega_b)\boldsymbol{X} \tag{2.10}$$

where $\mathbf{1}$ is column vector of ones.

The principal motivation for using the simplified matrices of Eq. (2.10) is that the mathematical treatment of the model becomes simpler at the expense of a moderately reduced flexibility [121, 247]. In particular, the two identities (used in similar derivations in Refs. [172, 174])

$$\omega_{rs} = \omega_b + x_{rs}(\omega_a - \omega_b)\ , \tag{2.11a}$$

$$\log \omega_{rs} = \log \omega_b + x_{rs}(\log \omega_a - \log \omega_b)\ , \tag{2.11b}$$

lead to a likelihood and a log-likelihood analogous to—but much simpler than—the ones appearing in Eqs. (2.7) and (2.8). They are associated with the score function

$$\begin{aligned} f_{\mathrm{GMGM}}(a_{ij}, \sigma_i, \sigma_j) &= a_{ij}\left[\log \omega_b + x_{\sigma_i\sigma_j}(\log \omega_a - \log \omega_b)\right] - \left[\omega_b + x_{\sigma_i\sigma_j}(\omega_a - \omega_b)\right] \\ &\sim x_{\sigma_i\sigma_j}[a_{ij}(\log \omega_a - \log \omega_b) - (\omega_a - \omega_b)] \\ &\sim x_{\sigma_i\ \sigma_j}\left[a_{ij} + \gamma\right] \end{aligned} \tag{2.12}$$

where $\gamma = (\omega_b - \omega_a)/(\log \omega_a - \log \omega_b) \in (-\infty, 0]$, a drastic simplification when contrasted with Eq. (2.9). In essence, the GMGM only cares about the type of a block pair. If a pair of nodes (i, j) is associated with a block pair of type a, then the global objective function is increased by a factor of $a_{ij} + \gamma$ (greater when $a_{ij} = 1$ than when $a_{ij} = 0$). If a pair of nodes (i, j) is associated with a block pair of type b, then it only has an indirect impact, by omission.

Combinatorial objective functions

The GMGM specialization of the SBM is interesting not only for its mathematical simplicity, but also because its pairwise score function can be obtained from a completely different perspective. As we have seen, the essence of the GMGM is its binary classification of block pairs ; it turns out that there are countless examples of MPE objective functions that rely on a similar dichotomy (see, for instance, Ref. [78] for a recent review). Their design is essentially the following. Some subsets or intersections of nodes are first identified as special. The MPE

[4] A parametrization where $\omega_b > \omega_a$ can be represented by an equivalent parametrization $(\boldsymbol{X}', \boldsymbol{\omega}')$, defined as $\boldsymbol{X}' = \mathbf{1}\mathbf{1}^T - \boldsymbol{X}$ with $\omega_a' = \omega_b$, and $\omega_b' = \omega_a$. The case $\omega_a = \omega_b$ is somewhat pathological, but it can be handled nonetheless, with $\boldsymbol{X} = \mathbf{1}\mathbf{1}^T$, $\omega_a' = \omega_a$ and $\omega_b' = \delta$, for any $\delta < \omega_a$.

objective function is then designed as to maximize the number of edges *within* or *to* these subsets. Finally, because there are often trivial maxima (e.g., place all nodes in the special subset), some constraints are added to avoid trivial optima.

The general mathematical construction closely parallels that of the GMGM. First, we designate special pairs of blocks, and encode the result in a binary matrix $\boldsymbol{X}$. We then assume, without loss of generality, that the number of edges within these blocks should be *maximized* by the target partition $\mathcal{B}^*$. This leads to the graphical objective function

$$\widetilde{H}(G|\mathcal{B}) = \sum_{i\leq j} a_{ij} x_{\sigma_i\sigma_j} \,. \tag{2.13}$$

Functions of the form of Eq. (2.13) are plagued by many trivial optima, since it is often possible to maximize $\widetilde{H}$ by placing all nodes in one or a few blocks. For instance, if $x_{rr} = 1$ for at least one r, then Eq. (2.13) is maximized by putting all nodes in block B_r—it is obvious that no mesoscopic information is contained in the resulting partition. In general, if Eq. (2.13) rewards placing many edges between some pair of blocks (B_r, B_s) via $x_{rs} = 1$, then it is possible to find good solutions simply by putting a lot of *nodes* in these blocks : The more nodes, the more edges, and therefore the better score. We discourage these uninformative solutions by introducing an additive *balance constraints* $h(\mathcal{B})$ that penalizes the objective function $\widetilde{H}$ for partitions that contain large blocks aligned with $\boldsymbol{X}$. Specifically, we use a quadratic constraint on the block sizes [172, 229]

$$h(\mathcal{B}) = 2\gamma \sum_{r,s} x_{rs} n_r n_s \,, \qquad \gamma < 0 \,, \tag{2.14}$$

where n_r is the size of block B_r, and where $|\gamma|$ controls the overall strength of the constraint h. Because the constraint appearing in Eq. (2.14) can be rewritten as

$$2\gamma \sum_{r,s} x_{rs} n_r n_s = 2\gamma \sum_{r,s} x_{rs} \left(\sum_{i=1}^{n} \delta_{\sigma_i r} \right) \left(\sum_{j=1}^{n} \delta_{\sigma_j s} \right) = \gamma \sum_{i\leq j} x_{\sigma_i\sigma_j}$$

where δ_{ab} is the Kronecker delta (equal to 1 if $a = b$ and to zero otherwise), the constrained version of Eq. (2.13) is equivalent to

$$H(G|\mathcal{B}) = \widetilde{H}(G|\mathcal{B}) + h(\mathcal{B}) = \sum_{i\leq j} x_{\sigma_i\sigma_j} [a_{ij} + \gamma] \,. \tag{2.15}$$

This balanced objective function is obviously associated with a pairwise score function equivalent to that of the GMGM [c.f. Eq. (2.12)]. Therefore, all objective functions formulated as an *edge count maximization* coupled with an additive *quadratic balance constraint* are equivalent to the GMGM. Furthermore, the strength of the balance constraint γ can be seen as a function of the parameters (ω_a, ω_b) of the corresponding GMGM : The greater the difference between ω_a and ω_b, the stronger the balance constraint.

The equivalence of the GMGM with combinatorial objective function has far reaching consequences, because many MPE methods are based on variation on these functions. A few well-known examples are : Balanced minimum cut, with $\boldsymbol{X} = \boldsymbol{I}$ where $\boldsymbol{I}$ is an identity matrix [173] ; approximative graph coloring with $\boldsymbol{X} = \mathbf{1}\mathbf{1}^\intercal - \boldsymbol{I}$ [56, 131] ; nonoverlapping core-peripheries (CP) under size constraints [125, 210] with, e.g.,

$$\boldsymbol{X}_{\text{CP1}} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \qquad \boldsymbol{X}_{\text{CP2}} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}, \qquad \boldsymbol{X}_{\text{MultiCP}} = \begin{pmatrix} 1 & 1 & \dots & 0 & 0 \\ 1 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 1 \\ 0 & 0 & \dots & 1 & 0 \end{pmatrix}.$$

If anything, these simple examples show that the GMGM and Eq. (2.15) can be used as an “objective function factory” of sort : For any choice of γ and q, there will be $2^{\binom{q}{2}+q}$ different binary symmetric matrices $\boldsymbol{X}$, and as many MPE objective functions. Those that are named and well studied are but a tiny fraction of the full spectrum of possibilities ; most will uncover exotic patterns that are mixtures of core-peripheries, cuts, coloring, hierarchies, etc.

2.6.2 Complete hierarchy

While the SBM and its GMGM are general enough to specialize to many well-known MPE methods, there are also numerous objective functions that cannot be written as in Eqs. (2.9) and (2.12)—e.g., modularity functions—, because they rely also on some *side-information* matrix $\boldsymbol{\lambda}$ absent from the pairwise scores of Eq. (2.9). The purpose of the present section is to expand on the classification of Sec. 2.6.1 to accommodate these functions.

Stochastic block model with side information

In the spirit of Ref. [116], we define a generalization of the Poisson SBM, whose likelihood is given by

$$\mathbb{P}(G|\mathcal{B},\boldsymbol{\omega},\boldsymbol{\Lambda}) = \prod_{i\leq j} \frac{(\omega_{\sigma_i\sigma_j}\lambda_{ij})^{a_{ij}}}{a_{ij}!} e^{-\omega_{\sigma_i\sigma_j}\lambda_{ij}} . \tag{2.16}$$

This (over-parametrized) version of the SBM combines mesoscopic information (via $\boldsymbol{\omega}$) with side information at the level of edges (via $\boldsymbol{\Lambda}$). It directly specializes to many well–known likelihoods, including the classical Poisson SBM (with $\boldsymbol{\Lambda} = \mathbf{1}\mathbf{1}^\intercal$), or the degree-corrected SBM of Ref. [116] (with $\boldsymbol{\Lambda} = \boldsymbol{k}\boldsymbol{k}^\intercal/2m$ where $\boldsymbol{k}$ is the vector of degrees).

As with its classical counterpart, one can find the most likely partition of the nodes of G by maximizing the logarithm of the likelihood (2.16) :

$$\log \mathbb{P} = \sum_{i\leq j} \left[a_{ij} \log \omega_{\sigma_i\sigma_j}\lambda_{ij} - \omega_{\sigma_i\sigma_j}\lambda_{ij} - \log a_{ij}! \right] .$$

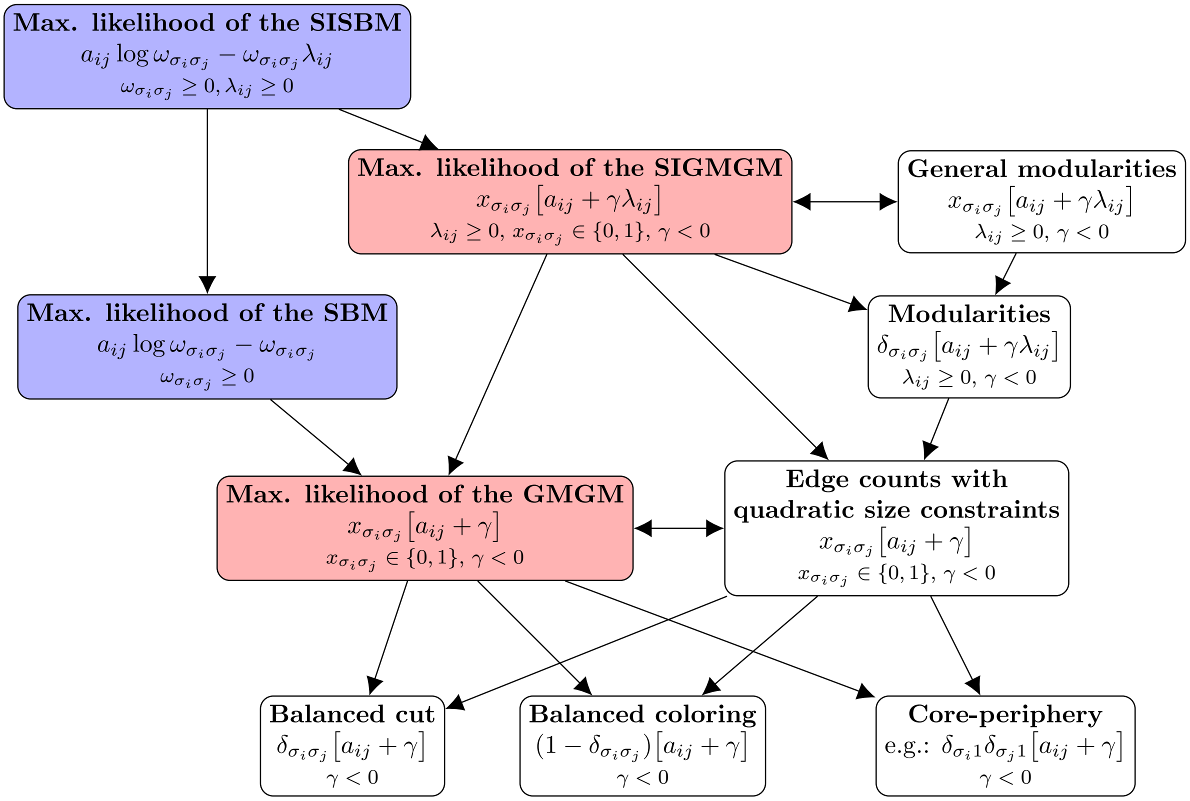

Max. likelihood of the SISBM
$a_{ij}\log\omega_{\sigma_i\sigma_j} - \omega_{\sigma_i\sigma_j}\lambda_{ij}$
$\omega_{\sigma_i\sigma_j} \geq 0, \lambda_{ij} \geq 0$
Max. likelihood of the SIGMGM
$x_{\sigma_i\sigma_j}\big[a_{ij} + \gamma\lambda_{ij}\big]$
$\lambda_{ij} \geq 0$, $x_{\sigma_i\sigma_j} \in \{0,1\}$, $\gamma < 0$
General modularities
$x_{\sigma_i\sigma_j}\big[a_{ij} + \gamma\lambda_{ij}\big]$
$\lambda_{ij} \geq 0$, $\gamma < 0$
Max. likelihood of the SBM
$a_{ij}\log\omega_{\sigma_i\sigma_j} - \omega_{\sigma_i\sigma_j}$
$\omega_{\sigma_i\sigma_j} \geq 0$
Modularities
$\delta_{\sigma_i\sigma_j}\big[a_{ij} + \gamma\lambda_{ij}\big]$
$\lambda_{ij} \geq 0$, $\gamma < 0$
Max. likelihood of the GMGM
$x_{\sigma_i\sigma_j}\big[a_{ij} + \gamma\big]$
$x_{\sigma_i\sigma_j} \in \{0,1\}$, $\gamma < 0$
Edge counts with quadratic size constraints
$x_{\sigma_i\sigma_j}\big[a_{ij} + \gamma\big]$
$x_{\sigma_i\sigma_j} \in \{0,1\}$, $\gamma < 0$
Balanced cut
$\delta_{\sigma_i\sigma_j}\big[a_{ij} + \gamma\big]$
$\gamma < 0$
Balanced coloring
$(1 - \delta_{\sigma_i\sigma_j})\big[a_{ij} + \gamma\big]$
$\gamma < 0$
Core-periphery
e.g.: $\delta_{\sigma_i 1}\delta_{\sigma_j 1}\big[a_{ij} + \gamma\big]$
$\gamma < 0$

FIGURE 2.3 – Partial hierarchy of objective functions. The pairwise score function of MPE methods are shown with the range of parameters below. Arrows denote specialization; doubled–sided arrows denote equivalence. Only the most direct arrows are drawn for the sake of clarity; specialization and equivalence are transitive operations. The abbreviations are : stochastic block model (SBM), with side information (SISBM); and general modular graph modular model (GMGM), with side information (SIGMGM). Functions derived from the perspective of statistical inference are colored in blue (general classification) and red (binary classification).

Therefore, the maximum likelihood formulation of the SBM with side information (hereafter : SISBM) is associated with the pairwise score function

$$f_{\text{SISBM}}(a_{ij},\lambda_{ij},\sigma_i,\sigma_j) \sim a_{ij}\log\omega_{\sigma_i\sigma_j} - \omega_{\sigma_i\sigma_j}\lambda_{ij}\,. \tag{2.17}$$

The likelihood is not useful in itself, because there are too many parameters for the amount of information encoded in $\boldsymbol{A}$. However, considering Eq. (2.17) not as a proper MPE method, but rather as the starting point of a general objective function hierarchy, it becomes a useful classification tool.

General modular graph model with side information

As with the classical SBM, it is possible to define a GMGM specialization of the SISBM. Following Sec. 2.6.1, the idea is again to classify all pairs of blocks according to their density category (via $\boldsymbol{X}$), and to re-use the identities appearing in Eq. (2.11) to rewrite the log-likelihood.

Model	λ_{ij}	Ref.
Configuration model (CM)	$k_i k_j/(2m)$	[176]
CM with resolution	$\zeta k_i k_j/(2m)$	[207]
Erdős-Rényi	ρ	[232]
Constant Potts model	ζ	[232]
Gravity model	$k_i k_j \phi(r_{ij})$	[71]

TABLE 2.1 – Examples of null models. In all cases, $\zeta > 0$ is a free parameter, $\rho \in [0,1]$ is the density of the network. The gravity model is included as an example of an exotic null model ; it is derived for spatially embedded network, with r_{ij} being the Euclidean distance between nodes i and j and ϕ some reference connection propensity in space.

The resulting likelihood is still over-parametrized because of $\boldsymbol{\Lambda}$, but much simpler than that of the general SISBM, since the connection matrices $\boldsymbol{\omega}$ are now restricted to the form of Eq. (2.10). It is easy to show that the pairwise score function is now

$$\begin{aligned} f_{\text{SIGMGM}}(a_{ij},\lambda_{ij},\sigma_i,\sigma_j) &= a_{ij}\left[\log\omega_b + x_{rs}(\log\omega_a - \log\omega_b)\right] - \left[\omega_b + x_{\sigma_i\sigma_j}(\omega_a - \omega_b)\right]\lambda_{ij} \\ &\sim x_{\sigma_i\sigma_j}\left[a_{ij}(\log\omega_a - \log\omega_b) - \lambda_{ij}(\omega_a - \omega_b)\right] \\ &\sim x_{\sigma_i\sigma_j}\left[a_{ij} + \gamma\lambda_{ij}\right] \end{aligned} \tag{2.18}$$

where $\gamma < 0$ is the same parameter as the one appearing in Eq. (2.12).

Modularity functions

One of the reasons why the GMGM specialization is useful is, again, that it can be derived from first principles in a completely different manner, this time from the point of view of the *modularity* [176]. In a nutshell, modularity is defined as the difference between the number of internal edges of a partition (edges that connect two nodes in the same block), and the *expected* number of internal edges for this partition, if the network were to be drawn from some null model. The idea behind modularity is to maximize the number of edges within blocks, while accounting for the edges that would have been there in the first place, just by pure chance (assuming some model for the network, see Table. 2.1).

Modularity is a graphical objective function, since it can be written as a sum over pairs of nodes [169, 207]. Writing the expected number of edges between the nodes i and j as λ_{ij} under the null model of choice, the modularity of a partition reads

$$H_{\text{Mod}}(\mathcal{B},\boldsymbol{\lambda},G) \propto \sum_{i\leq j}[a_{ij} - \lambda_{ij}]\delta_{\sigma_i\sigma_j} \,. \tag{2.19}$$

Importantly, the pairwise score function associated with the modularity is always given by

$$f_{\text{Mod}}(a_{ij},\lambda_{ij},\sigma_i,\sigma_j) = (a_{ij} - \lambda_{ij})\delta_{\sigma_i\sigma_j} \sim (a_{ij} + \gamma\widetilde{\lambda}_{ij})\delta_{\sigma_i\sigma_j} \,, \tag{2.20}$$

independent from the choice of null model (where $\gamma < 0$ and $\widetilde{\lambda}_{ij}$ is a rescaled connection probability under the null model).

A comparison with Eq. (2.18) reveals that the above score function—and therefore that of *any* modularity-type function—is in fact a specialization of the GMGM with side information, recovered by setting $\boldsymbol{X} = \boldsymbol{I}$, $\gamma = -1$, and by using the null model of the modularity as the side-information matrix $\boldsymbol{\Lambda}$. In other words, every modularity function is equivalent to some variant of the SIGMGM, where the null model is multiplied with the connection matrix $\boldsymbol{\omega}$ where $\boldsymbol{X}$ is simply the identity.

It is worth pointing out that using a flat null model (i.e., $\widetilde{\lambda}_{ij} = 1$) in Eq. (2.20) amounts to opting for a GMGM without side information, with $\boldsymbol{X} = \boldsymbol{I}$. Because the latter is associated with a pairwise score functions that is equivalent to edge counts coupled with quadratic balance constraints, it follows that *flat null models* act exactly like quadratic balance constraints. This correspondence explains the regularization properties of the ER null model investigated in Ref. [211, 232] (among others).

We note in closing that while the connection between the modularity and the GMGM is presented here for $\boldsymbol{X} = \boldsymbol{I}$, it is of course possible to define "modularities" associated with different matrices $\boldsymbol{X}$, in the spirit of the side information free equivalence. These modularities will be able to uncover any mixture of mesoscopic patterns reflected in $\boldsymbol{X}$.

2.7 Discussion

In this paper, we have shown that the maximum likelihood formulation of the SBM is perfectly equivalent to a number of standard mesoscopic pattern extraction (MPE) methods, upon appropriate specialization of its density matrix $\boldsymbol{\omega}$. Specifically, we have found that different classes of density matrices are associated with various classes of MPE algorithms, such as minimum cuts, modularities, core-periphery algorithm and combinations thereof. This has allowed us to delineate a hierarchy of MPE methods (Fig. 2.3), and to understand all methods as increasingly simplified SBMs. In doing so, we have shown that the SBM is universal with respect to mesoscopic pattern extraction with graphical functions—a conclusion that is complementary to the recent observation that the SBM is a universal network approximator [180].

Apart from a better understanding of MPE methods, in the light of the hierarchy of Fig. 2.3, there are a number of practical consequences to the fact that many of the MPE methods of network science are, after all, the SBM in disguise. Let us mention a few in closing.

First and foremost, these equivalences imply that the efficient maximizers (see, e.g., Refs. [193]) developed to tackle the hard problem of estimating $\mathcal{B}$ for the general SBM can be reused to solve more specific MPE problems that are also hard. This application of the equivalences

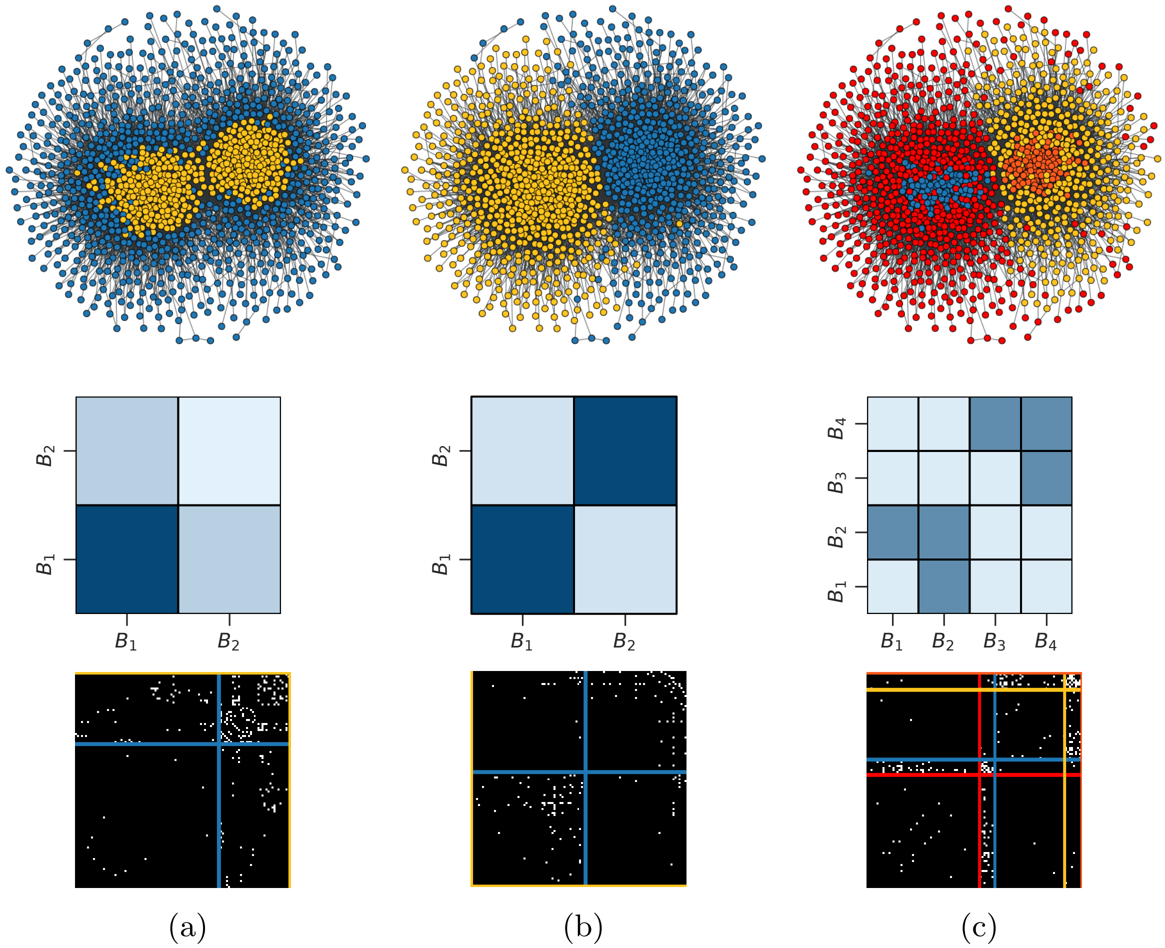


FIGURE 2.4 – Mesoscopic patterns learned from and imposed on a real complex network. All results are obtained on the `polblog` dataset, a directed network of hyperlinks between weblogs on US politics, recorded shortly after the 2004 presidential election. There are a total of 1 222 nodes (weblogs) and 16 714 edges. We use an undirected, self-loop free version of the network. All subfigures show (top) the network with nodes colored according to the identified partition, (center) a cartoon of the matrix $\boldsymbol{\omega}$ *imposed* for the equivalent SBM [darker shades of blue represent larger values of ω_{rs}], and (bottom) the adjacency matrix with the limit of blocks indicated as colored lines and edges as white dots. The optima of the objective functions are found via simulated annealing and greedy search [247]. (a) Natural partition of the network in $q = 2$ blocks, as found with the classical Bernoulli SBM via expectation–maximization (EM) on $\mathcal{B}$ and $\boldsymbol{\omega}$. In this case alone, $\boldsymbol{\omega}$ is learned and not imposed. (b) Balanced cut obtained with $\gamma \approx -25$ and the GMGM. The two blocks have size $\boldsymbol{n}^\intercal = [650, 572]$. A similar partition is identified by the modularity. (c) Double core-periphery found with $\gamma \approx -9$. The cores are of sizes 98 and 87 while their respective peripheries contain 396 and 641 nodes. There are only 3 703 edges between nodes of the peripheries (out of a maximum of 283 330 possible edges).

is direct : To optimize an MPE objective function, simply fix the matrix $\boldsymbol{\omega}$ (and $\boldsymbol{\Lambda}$ if there is side information) with some target mesoscopic pattern in mind, and run an SBM likelihood maximization procedure to uncover $\mathcal{B}$ (we have used this method to obtain the results of Fig. 2.4).

Second, as is also pointed out in Ref. [174] (for the special case of modularity), arbitrary MPE methods that are specializations of the SBM now stand on sounder statistical foundations, once their connection with the SBM is recognized. This is due to the fact that their free parameters—e.g. γ in Eqs. (2.12) and (2.18)—can be interpreted as functions of the connectivity matrix $\boldsymbol{\omega}$, thereby providing a statistically principled estimation procedure—expectation–maximization [56]—for otherwise arbitrary parameters.

Third, we can conclude that a number of hidden assumptions are built into popular MPE methods. In particular, they amount to fitting simplified SBMs by maximum likelihood, often with misspecified density matrices $\boldsymbol{\omega}$. Doing so is not a problem *per se*, because the goal of MPE is not always to find the most natural or most statistically robust decomposition of a network [216] ; it might instead be to of reveal different facets of the mesoscopic organization of a same network (see Fig. 2.4). However, one should bear in mind that using these MPE methods amounts to fitting an ill-defined model, with all the problems that this may bring about, such as missing the best description of a network, or preventing inference algorithms from converging at all [118, 119, 254].

Fourth, a knowledge of equivalences can help us better interpret the empirical outcomes of mesoscopic pattern extraction. Two algorithms may behave similarly on a set of networks not due to the robustness of the patterns therein, but because they share an equivalent notion of optimality. Hence, empirical studies that rely on many MPE algorithms—say, comparative analyses [82, 122]—can avoid being lured by what appears to be a strong consensus of many methods that actually implement the same notion of optimality.

Finally, the equivalences lead to a number of theoretical shortcuts. One, the consistency results derived for the SBM [1] apply directly to all MPE algorithms in the hierarchy, by specialization. The consistency of the SBM in most scaling regimes (and the existence of a *detectability limit* [56]) therefore extends to virtually every MPE algorithm studied thus far. Two, formal NP–hardness results can be extended to many MPE methods, using trivial reductions. For example, since it is known that modularity maximization is in NP–hard [31], the equivalence of modularity with the likelihood maximization of the GMGM specialization of the degree-corrected SBM [174] directly implies the NP–hardness of the latter, and therefore of the SBM. Three, the universality of the SBM suggests that there is an extension of the no free lunch Theorem of Ref. [189] to a more generalized notion of MPE problems—not just community detection.

Chapitre 3

Structure mésoscopique II : Cohérence en taille finie

Article original :

"Finite-size analysis of the detectability limit of the stochastic block model"

Jean-Gabriel Young [1], Patrick Desrosiers [1, 2], Laurent Hébert-Dufresne [3], Edward Laurence [1] et Louis J. Dubé [1]

[1] Département de Physique, de Génie Physique, et d'Optique
Université Laval, Québec (QC), G1V 0A6, Canada
[2] Centre de recherche de l'Institut universitaire en santé mentale de Québec
Québec (QC), G1J 2G3, Canada
[3] Santa Fe Institute, Santa Fe, New Mexico, 87501, USA

Référence : Phys. Rev. E **95**, 062304 (2017) [247]

‡ Ces sections sont reproduites directement de l'article orignal. Le contenu n'en a été modifié que pour se conformer au format exigé par la Faculté des études supérieures et postdoctorales de l'Université Laval.

3.1 Avant-propos

Dans le chapitre précédent, nous avons montré que le SBM peut être considéré comme un modèle quasi universel pour la structure mésoscopique des réseaux réels. Un corollaire évident de ce résultat est qu'une compréhension complète du SBM donne automatiquement une bonne compréhension globale de la détection des structures mésoscopiques par inférence statistique.

Dans ce chapitre, on explore une des conséquences de cette correspondance. On étudie la cohérence du SBM en taille finie, i.e., les conditions sous lesquelles le SBM peut identifier les blocs cachés (la partition plantée) d'un réseau fini, généré par le SBM lui-même. Un résultat de la correspondance entre modèles est la suivante : tout modèle doit nécessairement échouer là où le SBM échoue. Puisqu'on réussit à identifier des régions de paramètres où le SBM n'est pas cohérent, on conclut qu'il existe des réseaux avec une structure mésoscopique cachée *indétectable* pour toutes méthodes.

3.2 Résumé

Il a été récemment montré que le modèle stochastique par blocs (SBM) souffre d'un problème de détectabilité dans le régime creux en taille infinie. Par problème de détectabilité, on veut dire qu'il existe une région de l'espace des paramètres pour laquelle aucun algorithme ne peut identifier une partition corrélée avec la partition plantée (la partition utilisée pour générer les instances du modèle). Dans ce chapitre, on traite le problème de la détectabilité en taille *finie*, à l'aide d'une approche tirée de la théorie de l'information. On distingue deux concepts de détectabilité, soit la détectabilité moyenne du modèle, et la détectabilité individuelle d'une instance. Pour ces deux formulations de la détectabilité, on démontre qu'il existe de grandes classes de paramètres qui mènent à un même niveau de détectabilité moyenne et individuelle. Dans une étude de cas détaillée, on applique notre formalisme général à un cas spécial du SBM, ce qui nous permet d'obtenir des expressions très explicites avec peu de paramètres. Ce cas spécial demeure toutefois assez général, puisqu'il s'applique entres autres : au SBM symétrique, au modèle de la coloration plantée, et à plusieurs variations non-triviales du SBM. On conclut avec des annexes, dans lesquels on étudie l'effet de la corruption aléatoire des données sur la limite de détectabilité et la connexion entre notre démarche et une démonstration équivalente, via la théorie des matrices aléatoires. On relègue les preuves plus techniques au dernier annexe.

3.3 Abstract

It has been shown in recent years that the stochastic block model (SBM) is sometimes undetectable in the sparse limit, i.e., that no algorithm can identify a partition correlated with the

partition used to generate an instance, if the instance is sparse enough and infinitely large. In this contribution, we treat the finite case explicitly, using arguments drawn from information theory and statistics. We give a necessary condition for finite-size detectability in the general SBM. We then distinguish the concept of average detectability from the concept of instance-by-instance detectability and give explicit formulas for both definitions. Using these formulas, we prove that there exist large equivalence classes of parameters, where widely different network ensembles are equally detectable with respect to our definitions of detectability. In an extensive case study, we investigate the finite-size detectability of a simplified variant of the SBM, which encompasses a number of important models as special cases. These models include the symmetric SBM, the planted coloring model, and more exotic SBMs not previously studied. We conclude with three appendices, where we study the interplay of noise and detectability, establish a connection between our information-theoretic approach and random matrix theory, and provide proofs of some of the more technical results.

3.4 Introduction

Mesoscopic analysis methods [203] are among the most valuable tools available to applied network scientists and theorists alike. Their aim is to identify regularities in the structure of complex networks, thereby allowing for a better understanding of their function [76, 171, 203], their structure [195, 218], their evolution [97, 248], and of the dynamics they support [99, 165, 213]. Community detection is perhaps the best-known method of all [76, 203], but it is certainly not the only one of its kind [171]. It has been shown, for example, that the separation of nodes in a core and a periphery occurs in many empirical networks [28], and that this separation gives rise to more exotic mesoscopic patterns such as overlapping communities [244]. This is but an example—there exist multitudes of decompositions in structures other than communities that explain the shape of networks both clearly and succinctly [190].

The stochastic block model (SBM) has proven to be versatile and principled in uncovering these patterns [105, 106, 238]. According to this simple generative model, the nodes of a network are partitioned in blocks (the *planted partition*), and an edge connects two nodes with a probability that depends on the partition. The SBM can be used in any of two directions : Either to generate random networks with a planted mesoscopic structure [99, 165] or to infer the hidden mesoscopic organization of real complex networks, by fitting the model to network datasets [105, 190, 224]—perhaps its most useful application.

Stochastic block models offer a number of advantages over other mesoscopic pattern detection methods [171]. One, there is no requirement that nodes in a block be densely connected, meaning that blocks are much more general objects than communities. Two, the sound statistical principles underlying the SBM naturally solve many hard problems that arise in network mesoscopic analysis ; this includes the notoriously challenging problem of determining the optimal number of communities in a network [122, 177, 192], or of selecting among the many possible descriptions of a network [122, 196, 203].

Another advantage of the statistical formulation of the SBM is that one can rigorously investigate its limitations. It is now known, for example, that the SBM admits a *resolution limit* [192] akin to the limit that arises in modularity–based detection method [77]. The limitations that have attracted the most attention, however, are the *detectability limit* and the closely related concept of *consistency limit* [1]. The SBM is said to be detectable for some parameters if an algorithm can construct a partition correlated with the planted partition [1], using no information other than the structure of a single—infinitely large—instance of the model. It is said to be consistent if one can *exactly* recover the planted partition. Therefore, consistency begets detectability, but not the other way around. Understanding when and why consistency (or detectability) can be expected is important, since one cannot trust the partitions extracted by SBM if it operates in a regime where it is not consistent (or detectable) [1].

Due to rapid developments over the past few years, the locations of the boundaries between the different levels of detectability are now known for multiple variants of the SBM, in the limit of infinite network sizes. If the average degree scales at least logarithmically with the number of nodes, then the SBM is consistent [2, 23], unless the constant multiplicative factor is too small, in which case the SBM is then detectable, but not consistent. If the average degree scales slower than logarithmically, then the SBM is at risk of entering an *undetectable* phase where no information on the planted partition can be recovered from the network structure [56, 208]. This happens if the average degree is a sufficiently small constant independent of the number of nodes.

These asymptotic results are, without a doubt, extremely useful. Many efficient algorithms have been developed to extract information out of hardly consistent infinite instances [55, 56, 130, 149]. Striking connections between the SBM and other stochastic processes have been established in the quest to bound the undetectable regime from below [1, 2, 14, 159]. But real networks are not infinite objects. Thus, even though it has been observed that there is a good agreement between calculations carried out in the infinite-size limit and empirical results obtained on small networks [55], it is not immediately clear that the phenomenology of the infinite case carries over, unscathed, to the finite case.

In this paper, we explicitly investigate detectability in *finite* networks generated by the SBM. We understand detectability in the information-theoretic sense [14]; our analysis is therefore algorithm–independent, and yields the boundaries of the region of the parameter space where the planted partition is undetectable, even for an optimal algorithm (with possibly exponential running time).

[1] By *correlated*, it is meant that the two partitions are more similar than two randomly constructed partitions. Our choice of measure will be made explicit at a later stage.

The combination of this information-theoretic point of view with our finite-size analysis leads to new insights and results, which we organize as follows. We begin by formally introducing the SBM and the necessary background in Sec. 3.5. We use this section to briefly review important notions, including inference (Sec. 3.5.2), as well as the consistency and detectability of the infinite SBM (Sec. 3.5.3). In Sec. 3.6, we present a necessary condition for detectability, and show that it is always met, on average, by finite instances of the SBM. We then establish the existence of a large equivalence class with respect to this notion of average detectability. In Sec. 3.8, we introduce the related concept of η–detectability and investigate the complete detectability distribution, beyond its average. In Sec. 3.9, we apply the perfectly general framework of Secs. 3.6–3.8 to a constrained variant of the SBM : the general modular graph model of Ref. [121]. The results of this section hold for a broad range of models, since the general modular graphs encompass the symmetric SBM, the planted coloring model, and many other models as special cases. We gather concluding remarks and open problems in Sec. 3.10. A few appendices follow. In the first, we investigate the interplay between noise and our notion of average detectability (Appendix 3.11.1); in the second, we establish a connection between our framework and random matrix theory (Appendix 3.11.2); in the remaing appendices, we give the details of two technical proofs encountered in the main text.

3.5 Stochastic block model

3.5.1 Definition of the model

The stochastic block model is formally defined as follows : Begin by partitioning a set of n nodes in q blocks of fixed sizes $\boldsymbol{n} = (n_1, ..., n_q)$, with $n = \sum_{r=1}^{q} n_r$. Denote this partition by $\mathcal{B} = \{B_1, ..., B_q\}$, where B_r is the set of nodes in the r^{th} block. Then, connect the nodes in block B_r to the nodes in block B_s with probability p_{rs}. In other words, for each pair of nodes (v_i, v_j), set the element a_{ij} of the adjacency matrix $\boldsymbol{A}$ to 1 with probability $p_{\sigma(v_i)\sigma(v_j)}$ and to 0 otherwise, where $\sigma(v_i)$ is the block of v_i. Note that for the sake of clarity, we will obtain all of our results for simple graphs, where edges are undirected and self-loops (edges connecting a node to itself) are forbidden[2]. This implies that $p_{rs} = p_{sr}$ and that $a_{ii} = 0$.

We will think of this process as determining the outcome of a random variable, whose support is the set of all networks of n nodes. Due to the independence of edges, the probability (likelihood) of generating a particular network G is simply given by the product of $\binom{n}{2}$ Bernoulli random variables, i.e.,

$$\mathbb{P}(G|\mathcal{B}, \boldsymbol{P}) = \prod_{i<j} [1 - p_{\sigma(v_i)\sigma(v_j)}]^{1-a_{ij}} [p_{\sigma(v_i)\sigma(v_j)}]^{a_{ij}} \ , \tag{3.1}$$

[2] There is no obstacle to a generalization to the directed case (with or without self-loops).

where $\boldsymbol{P}$ is the $q \times q$ matrix of connection probabilities of element p_{rs} (sometimes called the affinity or density matrix), and $i < j$ is a shorthand for "$i,j:\ 1 \leq i < j \leq n$." It is easy to check that the probability in Eq. (3.1) is properly normalized over the set of all networks of n distinguishable nodes.

A useful alternative to Eq. (3.1) expresses the likelihood in terms of the number of edges between each pair of blocks (B_r, B_s) rather than as a function of the adjacency matrix [224]. Notice how the number of edges m_{rs} appearing between the sets of nodes B_r and B_s is at most equal to

$$m_{rs}^{\max} = \begin{cases} \binom{n_r}{2} & \text{if } r = s, \\ n_r n_s & \text{otherwise.} \end{cases} \tag{3.2}$$

Each of these $m_{rs}^{\max}$ edges exists with probability p_{rs}. This implies that m_{rs} is determined by the sum of $m_{rs}^{\max}$ Bernoulli trials of probability p_{rs}, i.e., that m_{rs} is a binomial variable of parameter p_{rs} and maximum $m_{rs}^{\max}$. The probability of generating a particular instance G can therefore be written equivalently as

$$\mathbb{P}(G|\mathcal{B},\boldsymbol{P}) = \prod_{r \leq s} (1 - p_{rs})^{m_{rs}^{\max} - m_{rs}} (p_{rs})^{m_{rs}} \,. \tag{3.3}$$

where $\{m_{rs}\}$ and $\{m_{rs}^{\max}\}$ are jointly determined by the partition $\mathcal{B}$ and the structure of G, and $r \leq s$ denotes "$r,s:\ 1 \leq r \leq s \leq q$."

Having a distribution over all networks of n nodes, one can then compute average values over the ensemble. For example, the average degree of node v_i is given by

$$\langle k_i \rangle = \sum_r p_{\sigma(v_i) r} (n_r - \delta_{\sigma(v_i) r}) \,, \tag{3.4}$$

where δ_{ij} is the Kronecker Delta. The expression correctly depends on the block $B_{\sigma}(v_i)$ of v_i; nodes in different blocks will, in general, have different average degree. Averaging over all nodes, one finds the average degree of the network

$$\langle k \rangle = \frac{2}{n} \sum_{r \leq s} m_{rs}^{\max} p_{rs} \,. \tag{3.5}$$

This global quantity determines the density of the SBM when $n \to \infty$. The SBM is said to be dense if $\langle k \rangle = O(n)$, i.e., if p_{rs} is a constant independent of n. It is said to be sparse if $\langle k \rangle = O(1)$, i.e., if $p_{rs} = c_{rs}/n$ goes to zero as n^{-1}. In the latter case, a node has a constant number of connections even in an infinitely large network—a feature found in most large scale real networks [170].

For finite instances, it will often be more useful to consider the average density directly. It is defined as the number of edges in G, normalized by the number of possible edges, i.e.,

$$\rho = \frac{\langle k \rangle}{n - 1} = \sum_{r \leq s} (m_{rs}^{\max} / m^{\max}) p_{rs} \equiv \sum_{r \leq s} \alpha_{rs} p_{rs} \,, \tag{3.6}$$

where $m^{\max} = \sum_{r\leq s} m_{rs}^{\max}$, and

$$\alpha_{rs} := m_{rs}^{\max} / m^{\max} \; . \tag{3.7}$$

The dense versus sparse terminology is then clearer : The density of sparse networks goes to zero as $O(n^{-1})$, while dense networks have a nonvanishing density $\rho = O(1)$.

3.5.2 Inference

Depending on the elements of $\boldsymbol{P}$, the SBM can generate instances reminiscent of real networks with, e.g., a community structure [171] ($p_{rr} > p_{rs}$) or a core-periphery organization [28] ($p_{11} > p_{12} > p_{22}$ and $p_{22} \sim 0$). However, the SBM really shines when it is used to infer the organization in blocks of the nodes of real complex networks—this was, after all, its original purpose [105].

To have inferred the mesoscopic structure of a network (with the SBM) essentially means that one has found the partition $\mathcal{B}^*$ and density matrix $\boldsymbol{P}^*$ that best describes it. In principle, it is a straightforward task, since one merely needs to (a) assign a likelihood $\mathbb{P}(\mathcal{B}, \boldsymbol{P}|G)$ to each pair of partition and parameters [see Eqs. (3.1)–(3.3)], then (b) search for the most likely pair $(\mathcal{B}^*, \boldsymbol{P}^*)$. Since there are exponentially many possible partitions, this sort of enumerative approach is of little practical use. Fortunately, multiple approximate and efficient inference tools have been proposed to circumvent this fundamental problem. They draw on ideas from various fields such as statistical physics [55, 56, 190], Bayesian statistics [224, 233], spectral theory [130, 149, 173, 174], and graph theory [43], to name a few, and they all produce accurate results in general.

3.5.3 Detectability and consistency

One could expect perfect recovery of the parameters and partition from most of these sophisticated algorithms. This is called the consistency property. It turns out, however, that all known inference algorithms for the SBM, as diverse as they might be, fail on this account. And their designs are not at fault, for there exists an explanation of this generalized failure.

Consider the density matrix of elements $p_{rs} = \rho \ \forall (r,s)$. It is clear that the block partition is irrelevant—the generated network cannot and will not encode the planted partition. Thus, no algorithm will be abe to differentiate the planted partition from other partitions. It is then natural to assume that inference will be hard or impossible if $p_{rs} = \rho + \epsilon_{rs}(n)$, where $\epsilon_{rs}(n)$ is a very small perturbation for networks of n nodes ; there is little difference between the uniform case and this perturbed case. In contrast, if the elements of $\boldsymbol{P}$ are widely different from one another, e.g., if $p_{rr} = 1$ and $p_{rs} = 0$ for $r \neq s$, then easy recovery should be expected.

Understanding where lies the transition between these qualitatively different regimes has been the subject of much recent research (see Ref. [1] for a recent survey). As a result, the regimes have been clearly separated as follows : (i) the undetectable regime, (ii) the detectable (but not consistent) regime and (iii) the consistent regime (and detectable). It has further been established that the scaling of ρ with respect to n determines which regime is reached, in the limit $n \to \infty$.

The SBM is said to be *strongly consistent* if its planted partition can be inferred perfectly, with a probability that goes to 1 as $n \to \infty$ (it is also said to be in the *exact recovery* phase). Another close but weaker definition of consistency asks that the probability of misclassifying a node goes to zero with $n \to \infty$ (the *weakly consistent* or *almost exact recovery* phase). These regimes prevail when $\boldsymbol{P}$ scales at least as fast as $\boldsymbol{P} = \log(n)\boldsymbol{C}/n$, where $\boldsymbol{C}$ is a $q \times q$ matrix of constants [2, 3, 23]. Predictably, most algorithms (e.g., those of Refs. [3, 43, 224]) work well in the exact recovery phase regime, since it is the easiest of all .

In the *detectable* (but not consistent) regime, exact recovery is no longer possible (the *partial recovery* phase). The reason is simple : through random fluctuations, some nodes that belong to, say, block B_1, end up connecting to other nodes as if they belonged to block B_2. They are thus systematically misclassified, no matter the choice of algorithms. This occurs whenever $\boldsymbol{P} = \boldsymbol{C}/n$, or $\boldsymbol{P} = f(n)\boldsymbol{C}/n$, with $f(n)$ a function of n that scales slower than $\log(n)$.

The discovery of the third regime—the *undetectable regime*—arguably rekindled the study of the fundamental limits of the SBM [56, 208]. In this regime, which occurs when $\boldsymbol{P} = \boldsymbol{C}/n$ and $\boldsymbol{C}$ is more or less uniform, it is impossible to detect a partition that is even correlated with the planted one. That is, one cannot classify nodes better than at random, and no information on the planted partition can be extracted. Thus, some parametrizations of the SBM are said to lie below the *detectability limit*. This limit was first investigated with informal arguments from statistical physics [55, 56, 121, 208, 257], and has since been rigorously formalized in Refs. [14, 160], among others.

There exist many efficient algorithms that are reliable close to the detectability limit ; noteworthy examples include belief propagation [55, 56, 255], and spectral algorithms based on the ordinary [130] and weighted [159] non backtracking matrix, as well as matrices of self-avoiding walks [149]. But when the number of blocks is too large, most of these algorithms are known to fail well above the information-theoretic threshold, i.e., the point where it can be proven that the partition is detectable given arbitrary computational power. It has been therefore conjectured in Ref. [55], that there exists multiple troublesome phases for inference : A truly undetectable regime, and a regime where detection is not achievable *efficiently*. In the latter, it is thought that one *can* find a good partition, but only by enumerating all partitions—a task of exponential complexity.

In this contribution, however, we will not focus on this so-called hard regime. As far as we are concerned, detectability will be understood in terms of information, i.e., we will delimit the boundaries of the information-theoretically undetectable regime.

3.6 Detectability of finite networks

Detectability and consistency are well-separated phases of the infinite stochastic block model. A minute perturbation to the parameters may potentially translate into widely different qualitative behaviors. The picture changes completely when one turns to finite instances of the model. Random fluctuations are not smoothed out by limits, and transitions are much less abrupt. We argue that, as a result, one has to account for the complete distribution of networks to properly quantify detectability, i.e., define detectability for *network instances* rather than parameters. This, in turn, commands a different approach that we now introduce.

3.6.1 Hypothesis test and the detectability limit

Consider a single network G, generated by the SBM with some planted partition $\mathcal{B}$ and matrix $\boldsymbol{P} = r\mathbf{1}\mathbf{1}^\intercal + \boldsymbol{\epsilon}$, where $\mathbf{1}\mathbf{1}^\intercal$ is a matrix of ones, r a constant, and $\boldsymbol{\epsilon}$ a matrix of (small) fluctuations. Suppose that the average density equals ρ, and consider a second density matrix $\rho\mathbf{1}\mathbf{1}^\intercal$ for which the block structure has no effect on the generative process. If an observer with *complete knowledge* of the generative process and its parameters cannot tell which density matrix, $\boldsymbol{P}$ or $\rho\mathbf{1}\mathbf{1}^\intercal$, is the most likely to have generated G, then it is clear that *this particular instance* does not encode the planted partition. As a result, it will be impossible to detect a partition correlated with the planted partition.

This idea can be translated into a mathematical statement by way of a likelihood test. For a SBM of average density ρ, call the ensemble of Erdős-Rényi graphs of density ρ the ensemble of *equivalent random networks*. Much like the SBM (see Sec. 3.5), its likelihood $Q(G|\rho)$ is given by the product of the density of $\binom{n}{2}$ independent and identically distributed Bernoulli variables, i.e.,

$$\mathbb{Q}(G|\rho) = \prod_{i<j} \rho^{a_{ij}} (1-\rho)^{a_{ij}} = \rho^{m} (1-\rho)^{m^{\max} - m} \,, \tag{3.8}$$

where $m := \sum_{r \leq s} m_{rs}$ is the total number of edges in G.

The condition is then the following : Given a network G generated by the SBM of average density ρ and density matrix $\boldsymbol{P}$, one can detect the planted partition $\mathcal{B}$ only if the SBM is more likely than its equivalent random ensemble of density ρ, i.e.,

$$\Lambda = \frac{\mathbb{P}(G|\mathcal{B}, \boldsymbol{P})}{\mathbb{Q}(G|\rho)} > 1 \,. \tag{3.9}$$

A similar condition has been used in Refs. [160] and [14] to pinpoint the location of the detectability limit in infinite and sparse instances of the SBM. Nothing forbids its application to the finite-size problem ; we will see shortly that it serves us well in the context of finite-size detectability.

3.6.2 Normalized log-likelihood ratio

The (equivalent) normalized log-likelihood ratio

$$\mathcal{L} := \frac{\log \Lambda}{m^{\max}} \tag{3.10}$$

will be more practical for our purpose. This simple transformation brings the line of separation between models from $\Lambda = 1$ to $\mathcal{L} = 0$, and prevents the resulting quantity from becoming too large. More importantly, it changes products into sums and allows for a simpler expression,

$$\mathcal{L} = \sum_{r \leq s} \left\{ \frac{m_{rs}}{m^{\max}} \log\left[\frac{p_{rs}(1-\rho)}{\rho(1-p_{rs})}\right] + \alpha_{rs} \log\left[\frac{1-p_{rs}}{1-\rho}\right] \right\}. \tag{3.11}$$

We will focus, for the remainder of this chapter, on the case where network instances G of n nodes are drawn from the SBM of parameters $(\mathcal{B}, \boldsymbol{P})$. In this context, $\mathcal{L}$ is a random variable whose support is the networks of n nodes with labeled nodes (see Fig. 3.1). Since $\boldsymbol{P}, \rho, \alpha$, and $m^{\max}$ are all parameters, $\mathcal{L}$ can also be seen as a weighted sum of binomial distributed random variables $m_{rs} \sim \text{Bin}(m_{rs}^{\max}, p_{rs})$, with a constant offset. Its average will be a prediction of the detectability for the ensemble (Sec. 3.7), and the probability $\Pr(\mathcal{L} < 0; \boldsymbol{P}, \boldsymbol{\alpha}, m^{\max})$ will give the fraction of instances that are undetectable for the selected parameters (Sec. 3.8).

3.6.3 Interpretation of $\mathcal{L}$: Information-theoretic bound and inference difficulty

Because likelihood ratio tests can be understood as quantifying the amount of evidence for a hypothesis (compared to a null hypothesis), there will be two interpretations of $\mathcal{L}$.

On the one hand, the condition $\mathcal{L} > 0$ will provide a lower bound on detectability ; if $\mathcal{L}(G, \mathcal{B}, \boldsymbol{P}) < 0$, then we can safely say that the instance G is information-theoretically undetectable. However, $\mathcal{L}(G, \mathcal{B}, \boldsymbol{P}) > 0$ does not necessarily mean that the instance is information-theoretically detectable. This is due to the fact that the condition $\mathcal{L} > 0$ is necessary but not sufficient, since we assume a complete knowledge of the generative process in calculating $\mathcal{L}$.

On the other hand, we will interpret $\mathcal{L}$ operationally as a measure of the difficulty of the inference problem (not in the computational sense). A large ratio of a hypothesis $\mathbb{H}$ to its null model $\mathbb{H}_0$ implies that the hypothesis is a much better explanation of the data than $\mathbb{H}_0$; therefore, $\mathcal{L}$ measures how easy it is to select between $\mathbb{P}$ and $\mathbb{Q}$, given full knowledge of the generative process, and inference algorithms will perform better when the ratio is larger. Many empirical results will validate this interpretation (see Sec. 3.9).

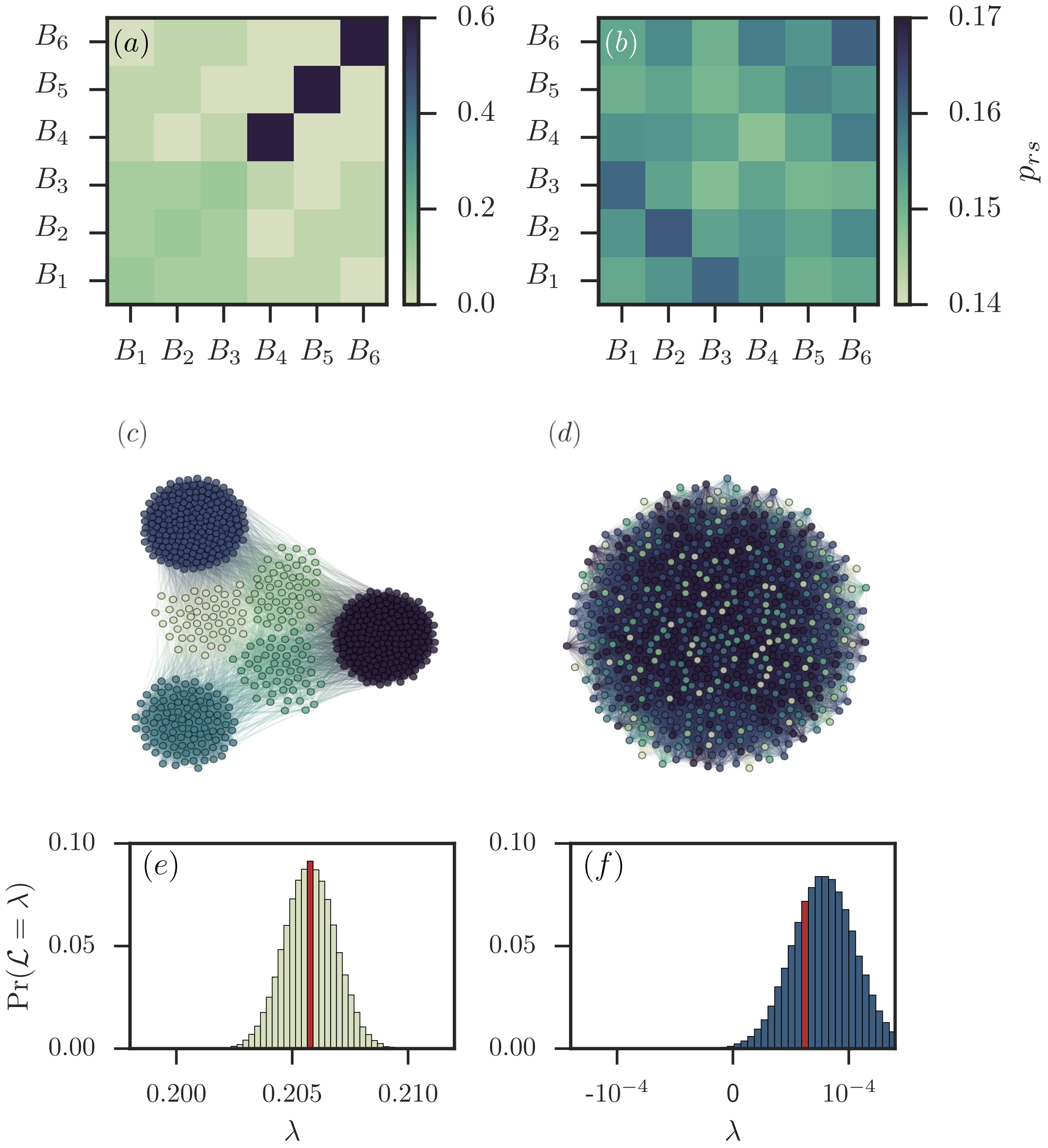


FIGURE 3.1 – Stochastic block model with (a, c, e) non uniform density matrix and (b, d, f) nearly uniform density matrix. (a, b) Density matrix of the two ensembles. Notice the difference in scale. (c, d) One instance of each ensemble, with $\boldsymbol{n} = [50, 50, 50, 100, 200, 200]$. Each color denotes a block [194]. (e, f) Empirical distribution of the normalized log-likelihood obtained from 100 000 samples of $\mathcal{L}$. The bins in which the instances (c, d) fall are colored in red. Notice that a negative log-likelihood ratio is associated with some instances in (f).

3.7 Average detectability

3.7.1 Average normalized log-likelihood

The average of a log-likelihood ratio is also known as the Kullback-Leibler (KL) divergence $D(\cdot||\cdot)$ of two hypotheses [48], i.e.,

$$\begin{aligned}\langle\mathcal{L}(\boldsymbol{\alpha},\boldsymbol{P})\rangle &= \sum_{\{G\}}\frac{\mathbb{P}(G|\mathcal{B},\boldsymbol{P})}{m^{\max}}\log\frac{\mathbb{P}(G|\mathcal{B},\boldsymbol{P})}{\mathbb{Q}(G|\rho)}\\ &= \frac{D(\mathbb{P}||\mathbb{Q})}{m^{\max}}\ ,\end{aligned} \tag{3.12}$$

where the sum runs over all n nodes networks. Since the KL divergence is always greater or equal to zero, with equality if and only if $\mathbb{P}=\mathbb{Q}$, and since $\mathcal{L}>0$ is only a necessary condition for detectability, the average $\langle\mathcal{L}\rangle$ will not be enough to conclude on detectability of the SBM, except for the case $\mathbb{P}=\mathbb{Q}$ as the only exception[3]. Results pertaining to $\langle\mathcal{L}\rangle$ will therefore be best interpreted in terms of inference difficulty.

However, even if the average log-likelihood ratio is always positive (assuming $\mathbb{P}\neq\mathbb{Q}$), it can be extremely close to zero for density matrix $\boldsymbol{P}$ "close" to $\rho\mathbf{1}\mathbf{1}^{\intercal}$ [Fig. 3.1 (f)]. In fact, as we will see in Sec. 3.8, $\langle\mathcal{L}(\boldsymbol{\alpha},\boldsymbol{P})\rangle\approx 0$ implies that there are instances for which $\mathcal{L}<0$. Therefore, whenever the average is small, we may also take it as a sign that the planted partition of some instances are truly undetectable.

3.7.2 Compact form

While Eq. (3.12) has a precise information-theoretic interpretation, there exists an equivalent form, both more compact and easier to handle analytically. It is given by

$$\langle\mathcal{L}(\boldsymbol{\alpha},\boldsymbol{P})\rangle = h(\rho)-\sum_{r\leq s}\alpha_{rs}h(p_{rs})\ , \tag{3.13}$$

where

$$h(p) = -(1-p)\log(1-p)-p\log(p) \tag{3.14}$$

is the binary entropy of $p\in[0,1]$. This expression can be obtained in a number of ways, the most direct of which is to take the average of Eq. (3.11) over all symmetric matrices $\boldsymbol{m}=(m_{11},m_{12},\ldots,m_{qq})$ with entries in $\mathbb{N}$ and upper bounds given by $\boldsymbol{m}^{\max}=(m_{11}^{\max},m_{12}^{\max},\ldots,m_{qq}^{\max})$. That is to say, we use the interpretation where $\mathcal{L}$ is a weighted sum of binomial distributed random variable, instead of the interpretation where it is a random variable over the networks of n nodes (see Sec. 3.6.2). The probability mass function associated with $\boldsymbol{m}$ is then $\Pr[\boldsymbol{m}]=\prod_{r\leq s}\Pr[m_{rs}]$, where $\Pr[m_{rs}]$ is the binomial distribution of parameter p_{rs} and upper bound $m_{rs}^{\max}$. Due to the linearity of expectations, it is straightforward to check that the

[3] $D(\mathbb{P}||\mathbb{Q})$ also goes to 0 at $\rho=0$, and a more careful scaling analysis is necessary to conclude on the detectability of sparse instances.

average of the first sum of Eq. (3.11) equals

$$\sum_{\boldsymbol{m}}\Pr[\boldsymbol{m}]\sum_{r\leq s}\frac{m_{rs}}{m^{\max}}\log\left[\frac{p_{rs}}{\rho}\frac{1-\rho}{1-p_{rs}}\right]=\sum_{r\leq s}\log\left[\frac{p_{rs}}{\rho}\frac{1-\rho}{1-p_{rs}}\right]\frac{m_{rs}^{\max}p_{rs}}{m^{\max}}\;.$$

Recalling Eq. (3.6), one then finds

$$\begin{aligned}\langle\mathcal{L}(\boldsymbol{\alpha},\boldsymbol{P})\rangle &= -\sum_{r\leq s}\alpha_{rs}\left[(1-p_{rs})\log(1-\rho)+p_{rs}\log\rho\right]\\ &\quad+\sum_{r\leq s}\alpha_{rs}[(1-p_{rs})\log(1-p_{rs})+p_{rs}\log p_{rs}]\\ &= h(\rho)-\sum_{r\leq s}\alpha_{rs}h(p_{rs})\;.\end{aligned}$$

where α_{rs} is defined in Eq. (3.7) with the normalization $\sum_{r\leq s}\alpha_{rs}=1$. Notice how this expression does not depend on $\mathcal{B}$ anymore. In this context, the only role of the planted partition is to fix the relative block sizes $\boldsymbol{\alpha}$. Thus, the average log-likelihood $\langle\mathcal{L}\rangle$ of two models with different planted partitions but identical $\boldsymbol{\alpha}$ is the same (up to a size-dependent constant).

With these two expressions for $\langle\mathcal{L}\rangle$ in hand [Eqs. (3.12) and (3.13)], we can now build an intuition for what the easiest and most difficult detectability problems might look like. The KL divergence is never negative, and Eq. (3.13) shows that the maximum of $\langle\mathcal{L}\rangle$ is $h(1/2)$; the average of the normalized log-likelihood is thus confined to the interval

$$0\leq\langle\mathcal{L}(\boldsymbol{\alpha},\boldsymbol{P})\rangle\leq h(1/2)\;. \tag{3.15}$$

An example of parameters that achieves the upper bound would be the SBM of density matrix $p_{11}=p_{22}=1$, $p_{12}=0$, with $\boldsymbol{n}=[n/2,n/2]$, i.e., the "ensemble" of disconnected $n/2$–cliques (which contains a single instance). An example of parameters that achieves the lower bound would be $\mathbb{P}=\mathbb{Q}$, but also $\rho\to 0$ [see Eq. (3.13)].

3.7.3 Equivalent stochastic block models

We now use Eq. (3.13) to uncover hidden connections between different regimes of the SBM. Notice how this expression induces equivalence classes in the parameter space of the model, with respect to $\langle\mathcal{L}\rangle$, i.e., subsets of parameters that all satisfy

$$\lambda=\langle\mathcal{L}(\boldsymbol{P},\boldsymbol{\alpha})\rangle\;, \tag{3.16}$$

where λ is a constant that defines the equivalence class.

In the next paragraphs, we will characterize these equivalence classes in two complementary ways. First, we will look for global transformations that preserve λ and map parameters $(\boldsymbol{\alpha},\boldsymbol{P})$ to some other—not necessarily close—pair of parameters $(\boldsymbol{\alpha}',\boldsymbol{P}')$. Provided that they satisfy a number of standard constraints, these transformations will be shown to correspond to the symmetry group of the set of *hypersurfaces* $\langle\mathcal{L}(\boldsymbol{\alpha},\boldsymbol{P})\rangle=\lambda$. Second, we will consider

Eq. (3.16) explicitly and use it to obtain an approximate hypersurface equation. This equation will be used in later sections to determine the location of the hypersurfaces that separate the parameter space of the SBM in different detectability phases.

Global transformations : The symmetry group of the SBM

We first look for the set of λ–preserving global transformations, i.e., all transformations $T(f_1, f_2)$ of the form

$$\boldsymbol{\alpha}' = f_1(\boldsymbol{\alpha}), \quad \boldsymbol{P}' = f_2(\boldsymbol{P}) \tag{3.17}$$

valid at every point of the parameter space. This is a broad definition and it must be restricted if we are to get anything useful out of it. Intuitively, we do not want these transformations to change the space on which they operate, so it is natural to ask that they be space-preserving. Under the (reasonable) constraint that these transformations are invertible as well, we can limit our search for λ–preserving transformations to the symmetry group of the parameter space.

We will be able to harness known results of geometry and algebra once the parameter space of the SBM is properly defined. This space is in fact the Cartesian product of two parameter spaces : The parameter space of $\boldsymbol{\alpha}$ and that of $\boldsymbol{P}$. Since there are $q^* = q(q+1)/2$ free parameters in both $\boldsymbol{\alpha}$ and $\boldsymbol{P}$, the complete space is of dimension $2q^* - 1$. It is the product of the q^*–dimensional hypercube—in which every point corresponds to a choice of $\boldsymbol{P}$—, and the $(q^* - 1)$–dimensional simplex—in which every point corresponds to a choice of $\boldsymbol{\alpha}$. The latter is a simplex due to the normalization $\sum_{r \leq s} \alpha_{rs} = (m^{\max})^{-1} \sum_{r \leq s} m_{rs}^{\max} = 1$.

Now, the symmetry group of the q^*–dimensional hypercube and that of the $(q^* - 1)$–dimensional regular simplex are well-known [49] : They are respectively the hyperoctahedral group B_{q^*} and the symmetric group S_{q^*}. Their action on $\boldsymbol{\alpha}$ and $\boldsymbol{P}$ can be described as

$$\begin{aligned} &\alpha_{rs} \mapsto \alpha'_{rs} = \alpha_{\pi(r,s)} \;, \\ &p_{rs} \mapsto p'_{rs} = \gamma_{rs} + (1 - 2\gamma_{rs}) p_{\omega(r,s)} \;, \end{aligned}$$

where $\gamma_{rs} = \{0,1\}$, and where both $\pi(r,s)$ and $\omega(r,s)$ are permutations of the indexes (r,s). While the symmetries of $\mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{P})$ are automatically symmetries of the parameters, the converse is not true. We therefore look for transformations T that satisfy

$$\langle \mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{P}) \rangle = \langle \mathcal{L}\big(f_1(\boldsymbol{\alpha}), f_2(\boldsymbol{P})\big) \rangle \;. \tag{3.18}$$

It turns out that this constraint is satisfied if and only if $\pi = \omega$ and $\gamma_{rs} = \gamma \; \forall (r,s)$, i.e., for transformations of the form

$$\alpha_{rs} \mapsto \alpha'_{rs} = \alpha_{\pi(r,s)} \;, \tag{3.19a}$$

$$p_{rs} \mapsto p'_{rs} = \gamma + (1 - 2\gamma) p_{\pi(r,s)} \;, \tag{3.19b}$$

with $\gamma = \{0,1\}$ (see Appendix 3.11.3 for a detailed proof). The permutation component of the symmetry is not to be confused with the symmetries generated by relabeling blocks : The latter only leads to $q!$ different symmetries, whereas the former correctly generates $q^*! \gg q!$ symmetries, or a total of $2q^*!$ symmetries once they are compounded with $p_{rs} \mapsto 1 - p_{rs}$. The symmetries come about because the ordering of summation of the terms $\alpha_{rs}h(p_{rs})$ in Eq. (3.13) does not matter, and both $h(\rho)$ and $h(p_{rs})$ are preserved when $p_{rs} \mapsto 1 - p_{rs}$.

As an example of symmetry, let us focus on the special transformation $p_{rs} \mapsto 1 - p_{rs}\ \forall(r,s)$ with $\pi(r,s) = (r,s)$, i.e., the only transformation that does not change the block structure of the model. Since networks generated by these parameters can be seen as complement of one another (i.e., an edge present in G is absent from G', and vice-versa), we may call this transformation the *graph complement* transformation. The fact that it preserves detectability can be understood on a more intuitive level with the following argument. Suppose that we are given an unlabeled network G generated by the SBM and that we are asked to confirm or infirm the hypothesis that it was, in fact, generated by the SBM. Even if nothing is known about the generative process, we can take the complement of the network—a trivial (and reversible) transformation. But this should not help our cause. After all, this transformation cannot enhance or worsen detectability since no information is added to or removed from G in the process. So we expect that λ be preserved, and it is. Because all other symmetries affect the block structure through a change of $\boldsymbol{\alpha}$, what the above result shows is that there is no other "information-preserving" transformation that can be applied to G without a prior knowledge of its planted partition.

Hypersurfaces and detectability regions

We now turn to the problem of finding compact and explicit formulas that describe the hypersurfaces of constant $\langle\mathcal{L}\rangle$ in the parameter space [see Eq. (3.16)]. In so doing we will have to be mindful of the fact that the scale $m^{\max}$ intervenes in the calculation, even though it is absent from our expression for $\langle\mathcal{L}\rangle$. This can be made explicit by rewriting Eq. (3.16) as $\langle\log\Lambda\rangle/m^{\max} = \widetilde{\lambda}$; it is easy to see that any given hypersurface will be comparatively closer to the region $\langle\mathcal{L}\rangle = 0$ in larger networks. We focus on the universal behavior of the hypersurfaces and remove all references to the scale of the problem by defining $\lambda := m^{\max}\widetilde{\lambda}$—predictions for real systems can be recovered by reverting to the correct scale.

While Eq. (3.16) is easily stated, it is not easily solved for, say, $\{p_{rs}\}$. The average normalized log-likelihood ratio involves a sum of logarithmic terms; the hypersurface equation is thus transcendental. To further complicate matters, there are $2q^* - 1 = q(q-1) - 1$ degrees of freedom and the number of free parameters grows quadratically with q. As a result, little can be said of truly general instances of the SBM—at least analytically. All is not hopeless, however, because there are approximation methods that work well when the number of free parameters is not too large. We sketch the idea here and apply it to a simpler variant of the

SBM in the case study of Sec. 3.9.

Expanding the binary entropy functions $h(p_{rs})$ around $p_{rs} = \rho\ \forall r \leq s$ drastically simplifies the hypersurface equation. Leaving the term $h(\rho)$ untouched, we find from Eq. (3.16)

$$\lambda = h(\rho) - \sum_{r\leq s} \alpha_{rs} \left[h(\rho) + \sum_{k=1}^{\infty} \frac{1}{k!} \frac{\partial^k h(x)}{\partial x^k}\bigg|_{x=\rho} (p_{rs} - \rho)^k \right] .$$

Due to the normalization of $\{\alpha_{rs}\}_{r\leq s}$, all terms in $h(\rho)$ cancel out, and the definition $\sum_{r\leq s} \alpha_{rs} p_{rs} = \rho$ allows us to eliminate the first order terms as well. We are therefore left with

$$2\lambda\rho(1-\rho) = \sum_{r\leq s} \alpha_{rs}(p_{rs} - \rho)^2 + \mathcal{O}[(p_{rs} - \rho)^3] \,, \tag{3.20}$$

where ρ is fixed and $(\boldsymbol{\alpha}, \boldsymbol{P})$ take on values constrained by both Eqs. (3.6) and (3.20). We then resort to a change of parameters and choose $\rho(\boldsymbol{P}, \boldsymbol{\alpha})$ as one of the parameters. Selecting the $q^* - 1$ other parameters Δ_{rs} such that $p_{rs} = \rho(\boldsymbol{P}, \boldsymbol{\alpha}) + \Delta_{rs}(\boldsymbol{P}, \boldsymbol{\alpha})$, we obtain the form

$$2\lambda\rho(1-\rho) = \sum_{r\leq s} \alpha_{rs}(\Delta_{rs})^2 \,. \tag{3.21}$$

Hypersurfaces are therefore ellipsoids when $p_{rs} \approx \rho\ \forall (r,s)$.

Besides the simplicity of Eq. (3.21), there are two additional arguments for dropping the higher order terms in Eq. (3.20). One, the series is invariant under the symmetry $p_{rs} \mapsto 1 - p_{rs}$ $\forall (r,s)$ only if we limit ourselves to the second-order expression : It is easily verified that

$$\frac{\partial^k h(x)}{\partial x^k}\bigg|_{x=\rho} (p_{rs} - \rho)^k = (-1)^k (k-2)! \left[\frac{1}{(\rho - 1)^{k-1}} - \frac{1}{(\rho)^{k-1}} \right] (p_{rs} - \rho)^k$$

is off by a sign for odd powers of k under the mapping $p_{rs} \mapsto 1 - p_{rs}$, which also implies $\rho \mapsto 1 - \rho$. Two, the true hypersurfaces enclose sets of parameters which are convex with respect to $\boldsymbol{P}$, and so does the hypersurface implicitly defined in Eq. (3.20). The convexity of the hypersurface follows from the fact that the sublevel set of a convex function encloses a convex set [30], and from the observation that $\langle \mathcal{L} \rangle$ is convex with respect to $\boldsymbol{P}$ [this is easy to show with Eq. (3.13) and the log-sum inequality ; see Appendix 3.11.4]. The convexity of the approximate level set is trivial to the second order, since it is an ellipsoid [Eq. (3.21)]. However, the approximate level set need not be convex when higher order terms are included. Together, these two observations tell us that while not exact, Eq. (3.20) captures the important *qualitative* features of the problem and that it is not necessarily true of approximate solutions with only a few extra terms.

3.8 Detectability distribution

In the previous section, we have computed the average $\langle \mathcal{L} \rangle$ and used it to obtain equivalence among the parameters, with respect to detectability. We have also shown that $\langle \mathcal{L} \rangle > 0$ for most parameters, i.e., that we could not use the necessary condition $\mathcal{L} > 0$ to conclude on the *undetectability* of the finite SBM, on average. As we will now argue, this conclusion must be further refined ; the full distribution of $\mathcal{L}$ leads to a more accurate picture of detectability.

3.8.1 The whole picture : η–detectability

Consider a parametrization $(\mathcal{B}, \rho\mathbf{1}\mathbf{1}^\intercal + \boldsymbol{\epsilon})$ of the SBM that yields $\langle\mathcal{L}\rangle \approx 0$. Turning to the distribution of $\mathcal{L}$ for this parametrization, one expects to find $\mathcal{L} < 0$ with non-zero probability (unless the distribution of $\mathcal{L}$ concentrates on $\mathcal{L} = 0$). Therefore, $\langle\mathcal{L}\rangle$ could be indicative of detectability for some *fraction* of the networks generated by the SBM.

Let us formalize this notion and introduce the concept of η–detectability. We will say that the ensemble of networks generated with the SBM of parameters $(\mathcal{B}, \boldsymbol{P})$ is η–detectable if

$$\Pr(\mathcal{L} < 0; \mathcal{B}, \boldsymbol{P}) = 1 - \eta \ . \tag{3.22}$$

That is, η gives the fraction of networks in the ensemble which evades the necessary condition for undetectability. If $\eta \to 0$, then detection is impossible, in the sense that most instances are best described by the null hypothesis $\mathbb{Q}$. If $\eta \to 1$, then most instances contain statistical evidence for $\mathcal{B}$; detection cannot be ruled out on the basis of the log-likelihood test.

We must compute the complete distribution or the cumulative distribution function of $\mathcal{L}$ to determine η. An exact result is out of reach since the outcome of $\mathcal{L}$ is determined by a weighted sum of independent binomial variables with non-identical distributions. In the following paragraphs, we give an approximate derivation based on the central limit theorem—it agrees extremely well with empirical results for all but the smallest networks.

3.8.2 Approximate equation for η

Equation (3.11) gives the normalized log-likelihood ratio as a sum of independent binomial random variables ; it can be written as

$$\mathcal{L} = \sum_{r\leq s} \frac{m_{rs}}{m^{\max}} x_{rs} + C \tag{3.23a}$$

where the constants x_{rs} and C are given by

$$x_{rs} = \log\left[\frac{p_{rs}}{\rho}\frac{1-\rho}{1-p_{rs}}\right] \ , \tag{3.23b}$$

$$C = \sum_{r\leq s} \alpha_{rs} \log\left[\frac{1-p_{rs}}{1-\rho}\right] \ , \tag{3.23c}$$

and where $m_{rs} \sim \mathrm{Bin}(p_{rs}, m_{rs}^{\max})$.

The central limit theorem (CLT) predicts that the distribution of an appropriately rescaled and centered transformation of $\mathcal{L}$ will converge to the normal distribution $\mathcal{N}(0,1)$ if the

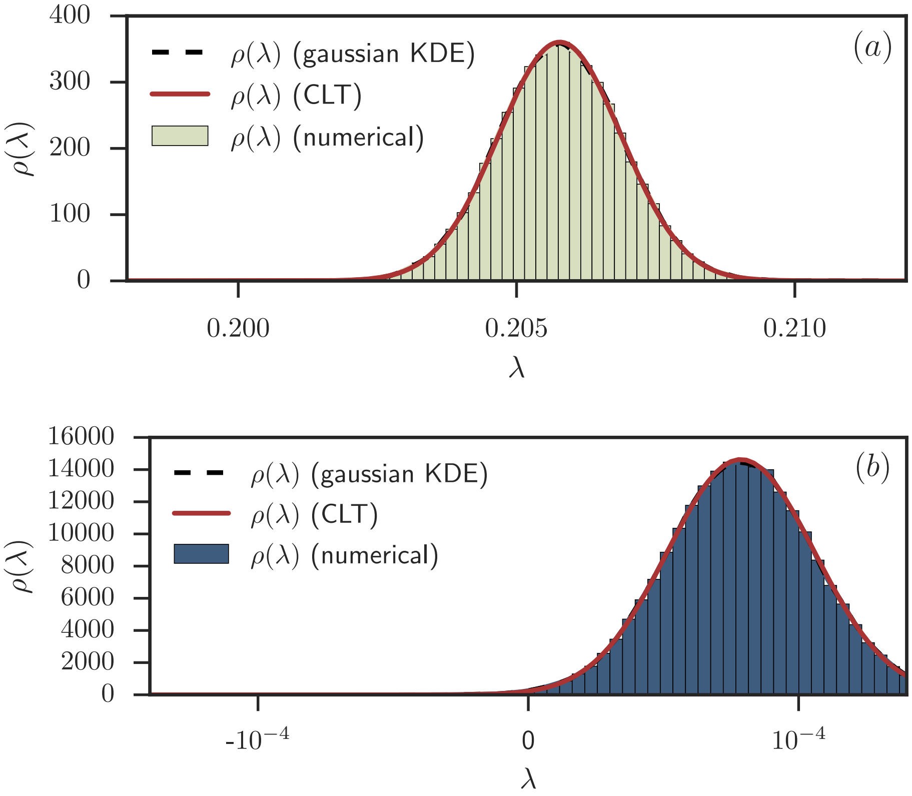


FIGURE 3.2 – Accuracy of the CLT approximation for the (a) non uniform and (b) nearly uniform SBM of Fig. 3.1. Both histograms aggregate 100 000 samples of $\mathcal{L}$. The prediction of the CLT is shown in red [see Eqs. (3.23b)–(3.23e)]. We plot for comparison the Gaussian kernel density estimate (KDE) of $\rho(\lambda)$ (dashed black line, hidden by the CLT curve). Equation (3.25) predicts $\eta_{(a)} = 1$ and $\eta_{(b)} = 0.981(2)$; for the sample shown, numerical estimates yield $\hat{\eta}_{(a)} = 1$ and $\hat{\eta}_{(b)} = 0.980(7)$.

number of summed random variables $q^* = q(q+1)/2$ goes to infinity. In the finite case, q^* obviously violates the conditions of the CLT, but it nonetheless offers a good approximation of the distribution of $\mathcal{L}$ (see Fig. 3.2).

To apply the CLT, we first define the centered and normalized variable $Z = (\mathcal{L} - C - \mu_{q^*})/S_{q^*}$, where

$$\begin{aligned} S_{q^*}^2 &= \sum_{r\leq s}\left[\left\langle \left(\frac{x_{rs}m_{rs}}{m^{\max}}\right)^2\right\rangle - \left\langle \left(\frac{x_{rs}m_{rs}}{m^{\max}}\right)\right\rangle^2\right] \\ &= \sum_{r\leq s}\frac{\alpha_{rs}}{m^{\max}}\,p_{rs}(1-p_{rs})x_{rs}^2 \end{aligned} \tag{3.23d}$$

is the sum of the variances of the q^* scaled binomial variables $x_{rs}m_{rs}/m_{rs}^{\max}$, and where

$$\mu_{q^*} = \sum_{r\leq s}\left\langle \frac{x_{rs}}{m^{\max}}m_{rs}\right\rangle = \sum_{r\leq s}\alpha_{rs}p_{rs}x_{rs} \equiv h(\rho) - \sum_{r\leq s}\alpha_{rs}h(p_{rs}) - C \tag{3.23e}$$

is the sum of their means [we have used Eq. (3.13) in the last step]. The CLT then tells us that $Z \sim \mathcal{N}(0,1)$, approximately.

Recall that the cumulative distribution function of a normal random variable can be expressed in terms of the error function as

$$\Pr(Z < z) = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{z}{\sqrt{2}}\right)\right] . \tag{3.24}$$

Now, assuming that Z is indeed normally distributed we can use the fact that $\Pr(\mathcal{L} < 0)$ is equivalent to $\Pr[Z < -(C + \mu_{q^*})/S_{q^*}]$ to compute η. Writing $\mu_{q^*} + C$ as $\langle \mathcal{L} \rangle$ [see Eq. (3.23e)], we find

$$\eta \approx \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{\langle \mathcal{L} \rangle}{\sqrt{2} S_{q^*}}\right)\right] , \tag{3.25}$$

i.e., an (approximate) equation in closed form for η.

Crucially, Eq. (3.25) predicts that η can never be smaller than $1/2$. This comes about because (i) $\langle \mathcal{L} \rangle > 0$ and (ii) S_{q^*} is a sum of variances, i.e., a positive quantity. There are therefore two possible limits which will yield $\langle \mathcal{L} \rangle / S_{q^*} \approx 0$ and $\eta = 1/2$: Either $\langle \mathcal{L} \rangle = 0$ or $S_{q^*} \gg 0$. Some care must be exerted in analyzing the case $\langle \mathcal{L} \rangle = 0$; Eqs. (3.11) and (3.12) tell us that the distribution of $\mathcal{L}$ is concentrated on 0 when its average is exactly equal to 0. We conclude that $\eta = 1/2$ is never reached but only approached asymptotically, for parameters that yield $\langle \mathcal{L} \rangle = \varepsilon$, with ε small but different from zero. The consequence of $\eta \geq 1/2$ is that at most half of the instances of the SBM can be declared undetectable on the account of the condition $\mathcal{L} < 0$.

3.8.3 Relation between average detectability and η–detectability

We can immediately reach a few conclusions on the interplay between the notions of average and η–detectability. First, the symmetries of $\langle \mathcal{L} \rangle$, (see Sec. 3.7.3) translates into symmetries for η. To see this, first notice that S_{q^*} is conserved under the mapping $p_{rs} \mapsto 1 - p_{rs}$

$$\begin{aligned} [x_{rs}(p_{rs},\rho)]^2 &\mapsto [-x_{rs}(1 - p_{rs}, 1 - \rho)]^2 , \\ p_{rs}(1 - p_{rs}) &\mapsto (1 - p_{rs})p_{rs} . \end{aligned}$$

and that a permutation of the indexes $\pi(r,s)$ only changes the order of summation of the terms of S_{q^*}. Second, hypersurfaces of constant average detectability need not be hypersurfaces of constant η–detectability.

To investigate this second important aspect of the connection between average detectability and η–detectability, let us further approximate Eq. (3.25). The MacLaurin series of the error function is, to the first order,

$$\begin{aligned}\eta &= \frac{1}{2}\left\{1+\frac{2}{\sqrt{\pi}}\left[\frac{\langle\mathcal{L}\rangle}{S_{q^*}} - \mathcal{O}\big(\langle\mathcal{L}\rangle^3/S_{q^*}^3\big)\right]\right\}\,,\\ &\approx \frac{1}{\sqrt{2\pi}}\frac{\langle\mathcal{L}\rangle}{S_{q^*}}+\frac{1}{2}\,. \end{aligned} \tag{3.26}$$

This is a reasonably accurate calculation of η when $\langle\mathcal{L}\rangle$ is small, i.e., close to the *average* undetectable regime. (Recall that we do not allow diverging S_{q^*} for the reasons stated in Sec. 3.8.2). It then becomes clear that on the hypersurfaces where $\langle\mathcal{L}\rangle = \lambda$ is constant (and close to 0),

$$\sqrt{2\pi}\left(\eta-\frac{1}{2}\right)S_{q^*} = \lambda\,, \tag{3.27}$$

is conserved rather than η itself. Equation (3.27) embodies a trade-off between accuracy (η) and variance (S_{q^*}) : In the regions of the hypersurface of constant $\langle\mathcal{L}\rangle$ where the variance is large, η must be comparatively small, and vice-versa.

3.8.4 1–detectability

Now, turning to the complementary case where $\langle\mathcal{L}\rangle$—and consequently η—is close to its maximum, we obtain a simple criterion for 1–detectability based on the asymptotic behavior of $\mathrm{erf}(x)$. It is reasonable to define a (small) threshold T beyond which $\mathrm{erf}(x > T) = 1$ for all practical purposes. The error function goes asymptotically to 1 with large values of its argument, but reaches its maximum of $\mathrm{erf}(x) = 1$ very quickly, so quickly, in fact, that $\mathrm{erf}(5)$ is numerically equal to 1 to the 10th decimal place.

Asking that the argument of $\mathrm{erf}(x)$ in Eq. (3.25) be greater than this practical threshold, we obtain the inequality

$$\langle\mathcal{L}\rangle \geq \sqrt{2}TS_{q^*} \tag{3.28}$$

for 1–detectability. Whenever the inequality holds, the associated ensemble is 1–detectable with a tolerance threshold T, i.e., we can say that for all practical purposes, there are no instances of the SBM that are necessarily [4] undetectable.

3.9 Case study : General Modular Graphs

The stochastic block model encompasses quite a few well-known models as special cases ; noteworthy examples include the *planted partition model* [43, 114], the closely related *symmetric SBM* (SSBM) [2, 56, 162], the *core-periphery model* [28], and many more. These simplified models are important for two reasons. One, they are good abstractions of structural patterns found in real networks, and a complete understanding of their behavior with respect to detectability is therefore crucial. Two, they are simple enough to lend themselves to a thorough

[4] Since $\mathcal{L} > 0$ is not sufficient for detectability, some instances could still be undetectable.

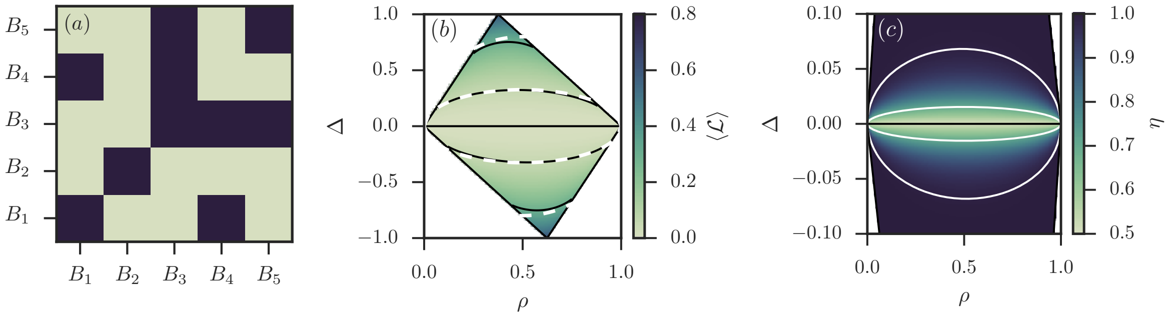


FIGURE 3.3 – Detectability in the general modular graph model. All figures use the same indicator matrix $\boldsymbol{W}$ [panel (a)] and the size vector $\boldsymbol{n} = [10,30,20,20,20]$ ($n = 100$ nodes). (a) Example of density matrix $\boldsymbol{P}$ allowed in the GMGM. Dark squares indicate block pairs of the inner type and light squares indicate pairs of the outer type. (b) Average detectability in the density space of the GMGM. Both the numerical solution of $\langle \mathcal{L} \rangle = \lambda$ (solid black line) and the prediction of Eq. (3.34) (dashed white line) are shown, for $\lambda = 0.05$ and 0.3. (c) $\eta(\rho, \Delta; \beta)$ in the density space of the GMGM ; notice the change of Δ–axis. Solid white lines are curves where $\eta(\rho, \Delta; \beta) = \eta^*$, with $\eta^* = 0.7$ (central curve) and $\eta^* = 0.99$ (outer curve). Equation (3.25) is used to compute both η and η^*.

analysis ; this contrasts with the general case, where simple analytical expressions are hard to come by.

In the paragraphs that follow, we investigate the *general modular graph model* (GMGM) [121], a mathematically simple, yet phenomenologically rich simplified model. Thanks to its simpler parametrization, we obtain easily interpretable versions of the expressions and results derived in Secs. 3.6–3.8.

3.9.1 Parametrization of general modular graphs

The GMGM can be seen as a constrained version of the SBM, in which *pairs* of blocks assume one of two roles : Inner or outer pairs. If a pair of blocks (B_r, B_s) is of the “inner type”, then one sets $p_{rs} = p_{\text{in}}$. If a pair of blocks (B_r, B_s) is of the “outer type”, then one sets $p_{rs} = p_{\text{out}}$. The resulting density matrices can therefore be expressed as

$$\boldsymbol{P} = (p_{\text{in}} - p_{\text{out}})\boldsymbol{W} + p_{\text{out}}\mathbf{1}\mathbf{1}^\intercal \,, \tag{3.29}$$

where $\boldsymbol{W}$ is a $q \times q$ indicator matrix [$w_{rs} = 1$ if (B_r, B_s) is an inner pair], and where $\mathbf{1}$ is a length q vector of ones. A non-trivial example of a density matrix of this form is shown in Fig. 3.3 (a). The figure is meant to illustrate just how diverse the networks generated by the GMGM may be, but it is also important to note that the results of this section apply to *any* ensemble whose density matrix can be written as in Eq. (3.29). This includes, for example, the q–block SSBM, a special case of the GMGM obtained by setting $\boldsymbol{W} = \boldsymbol{I}_q$ and $\{n_r = n/q\}_{r=1,..,q}$ (see Ref. [1] for a discussion of the SSBM).

While the parametrization in terms of p_{in} and p_{out} is simple, we will prefer an arguably more convoluted parametrization which is also more revealing of the natural symmetries of the GMGM (in line with the transformation proposed in Sec. 3.7.3). The first natural parameter is the average density, which can be computed from Eqs. (3.6) and (3.29) and which equals

$$\begin{aligned}\rho &= \sum_{r\leq s} \alpha_{rs}[w_{rs}p_{\text{in}} + (1-w_{rs})p_{\text{out}}]\ ,\\ &= \beta p_{\text{in}} + (1-\beta)p_{\text{out}}\ ,\end{aligned} \tag{3.30a}$$

where $\beta := \sum_{r\leq s}\alpha_{rs}w_{rs}$ is the fraction of *potential* edges that falls between block pairs of the inner type. The second natural parameter is simply the difference

$$\Delta = p_{\text{in}} - p_{\text{out}}\ . \tag{3.30b}$$

The absolute value of Δ quantifies the distance between the parameters of the GMGM and that of the equivalent random ensemble; its sign tells us which type of pairs is more densely connected. In this natural parametrization, the density matrix takes on the form $\boldsymbol{P} = \rho\mathbf{1}\mathbf{1}^\intercal + \Delta(1-\beta)\boldsymbol{W}$, i.e., a uniform matrix of ρ with perturbation proportional to $\Delta(1-\beta)$ for the inner pairs. It might appear that we have increased the complexity of the model description, since the additional parameter β now appears in the definition of the density matrix. It is, however, not the case, because we could consider the combined parameter $\widetilde{\Delta} = \Delta(1-\beta)$. Therefore, Eqs. (3.30a) and (3.30b), together with $\boldsymbol{W}$ and $\boldsymbol{n}$, suffice to unambiguously parametrize the model.

3.9.2 Average detectability of general modular graphs

The average normalized log-likelihood ratio $\langle\mathcal{L}\rangle$ is tremendously simplified in the natural parametrization of the GMGM; it is straightforward to show that the ratio takes on the compact (and symmetric) form

$$\langle\mathcal{L}(\rho,\Delta;\beta)\rangle = \beta\Big\{h(\rho) - h\big[\rho + (1-\beta)\Delta\big]\Big\} + (1-\beta)\Big\{h(\rho) - h\big[\rho - \beta\Delta\big]\Big\}\ , \tag{3.31}$$

by using $p_{rs} = w_{rs}p_{\text{in}} + (1-w_{rs})p_{\text{out}}$ together with the inverse of Eqs. (3.30a) and (3.30b),

$$p_{\text{in}} = \rho + (1-\beta)\Delta\ , \tag{3.32a}$$

$$p_{\text{out}} = \rho - \beta\Delta\ . \tag{3.32b}$$

In Fig. 3.3 (b), we plot $\langle\mathcal{L}(\rho,\Delta;\beta)\rangle$ in the (ρ,Δ) space—hereafter the density space—for the indicator matrix $\boldsymbol{W}$ shown in Fig. 3.3 (a) (and unequal block sizes, see caption). Unsurprisingly, $\langle\mathcal{L}\rangle$ is largest when the block types are clearly separated from one another, i.e., when $|\Delta|$ is the largest. Notice, however, how large separations are *not* achievable for dense or sparse networks. This is due to the fact that not all (ρ,Δ) pairs map to probabilities $(p_{\text{in}}, p_{\text{out}})$ in

$[0,1]$. The region of the density space that *does* yield probabilities is the interior of the quadrilateral whose vertices are, in (ρ,Δ) coordinates : $(0,0),(\beta,1),(1,0),(1-\beta,-1)$. Changing the value of β skews this accessible region and, presumably, the functions that are defined on it, such as $\langle\mathcal{L}(\rho,\Delta;\beta)\rangle$.

We also show on Fig. 3.3 (b) two members of the level set defined by $\langle\mathcal{L}(\rho,\Delta;\beta)\rangle=\lambda$. As mentioned previously, the exact functional form of this family of hypersurfaces (here simply curves) seems elusive, but an approximate solution is available. Using the method highlighted in Sec. 3.7, we find, to the second order,

$$\begin{aligned}2\lambda\rho(1-\rho) &\approx \sum_{r\leq s}\alpha_{rs}(p_{rs}-\rho)^2\\ &= \beta[(1-\beta)\Delta]^2+(1-\beta)(\beta\Delta)^2\,. \end{aligned} \tag{3.33}$$

Equation (3.33) fixes the relative value of all parameters on the line where $\langle\mathcal{L}\rangle=\lambda$. Solving for Δ, we find

$$\Delta^*(\rho;\lambda,\beta)=\pm\sqrt{2\lambda\frac{\rho(1-\rho)}{\beta(1-\beta)}}\,, \tag{3.34}$$

also shown on Fig. 3.3 (b) for comparison.

Figure 3.3 highlights the accuracy of our approximation when λ is small. But it also highlights its inaccuracy when λ is large ; $\lambda\gg 1$ forces $\Delta^*(\rho;\lambda,\beta)$ to pass through a region where $\Delta^*\approx 1$, i.e., a region where the omitted terms on the right-hand-side of Eq. (3.33) contribute heavily. Fortunately, this is not so problematic, since most detectability related phenomena—phase transitions, undetectable instances, etc.—arise near $\Delta=0$, i.e., where the approximation works.

3.9.3 η–detectability of general modular graphs

While $\langle\mathcal{L}(\rho,\Delta;\beta)\rangle$ takes on a particularly compact form once we substitute $\{p_{rs}\}$ by the natural parameters of the GMGM, the same cannot be said of $\eta(\rho,\Delta;\beta,n)$. Some analytical progress can be made by, e.g., noticing that only two types of terms are involved in the calculation of S_{q^*}, but, ultimately, the resulting expression is no more useful than the simple Eqs. (3.25) and (3.26). We will, therefore, omit the calculation of η.

In Fig. 3.3 (c) we plot $\eta(\rho,\Delta;\beta,n)$ in the density space [using Eq. (3.25)]. We also display the numerical solutions of $\eta(\rho,\Delta;\beta,n)=\eta^*$ for two values of η^*. The figure highlights just how quickly η goes to 1 as a function of Δ, even for the fairly small system sizes considered : We find that $\eta\geq 0.99$ for *any* value of ρ, as soon as $\Delta>0.06$. The condition in Eq. (3.9) is therefore a *weak* one. It allows us to determine that some parameters are overwhelmingly undetectable, but only when Δ is very close to 0.

Figure 3.3 also shows how increases in variance translate into decreases in accuracy [see Eq. (3.27)] : Following a line of constant (and relatively small) Δ, one can see that η is minimized close to $\rho = 1/2$, i.e., near the maximum of variance. This is characteristic of many parametrizations of the SBM and GMGM; it turns out that, for fixed n, impossible detection problems are not confined to vanishing densities. In fact, values of ρ closer to $1/2$ are associated with a comparatively larger interval of Δ for which detection is impossible.

3.9.4 Symmetries of general modular graphs

In Secs. 3.7 and 3.8, we have proved that there are $2q^*!$ transformations that preserve $\langle \mathcal{L}(\rho,\Delta;\beta)\rangle$ and $\eta(\rho,\Delta;\beta,n)$. We could therefore go about computing the symmetries of the GMGM by listing all of these transformations in terms of (ρ,Δ,β). But since there are only three free parameters in the GMGM, we can also choose an alternative route and directly solve $\langle \mathcal{L}(\rho,\Delta;\beta)\rangle = \langle \mathcal{L}(a_1\rho + b_1, a_2\Delta + b_2; a_3\beta + b_3)\rangle$ by, e.g., obtaining a linear system from the Taylor series of $\langle \mathcal{L}(\rho,\Delta;\beta)\rangle$. This simpler approach yields the following set of λ–preserving transformations for the model :

$$(\rho,\Delta,\beta) \mapsto (\rho,\Delta,\beta)\,, \tag{3.35a}$$

$$(\rho,\Delta,\beta) \mapsto (\rho,-\Delta,1-\beta)\,, \tag{3.35b}$$

$$(\rho,\Delta,\beta) \mapsto (1-\rho,\Delta,1-\beta)\,, \tag{3.35c}$$

$$(\rho,\Delta,\beta) \mapsto (1-\rho,-\Delta,\beta)\,. \tag{3.35d}$$

It is straightforward to check that these transformations form a group, whose product is the composition of two transformations. A Cayley table reveals that the group is isomorphic to the Klein four-group $Z_2 \times Z_2$.

One immediately notices a large gap between the number of symmetries predicted by the calculations of Sec. 3.7.3 ($2q^*!$) and the number of symmetries appearing in Eq. (3.35) (4, independent of q). The gap is explained by the fact that every symmetry of the general SBM maps onto one of the four transformations listed in Eq. (3.35) [5] A sizable fraction of the symmetries reduce to Eq (3.35a), since permutations $\pi(r,s)$ cannot modify the natural parameters of the GMGM : The type of block pair (B_r, B_s)—characterized by p_{rs}—is permuted alongside its share of potential edges α_{rs}. Another important fraction of the symmetries is accounted for by the "graph complement transformation" : Any transformation $\boldsymbol{P} = \mathbf{1}\mathbf{1}^\intercal - \boldsymbol{P}$ plus a permutation reduces to Eq. (3.35d). This leaves two symmetries, which happen to be artifacts of our choice of nomenclature. To see this, *rename* pair types, i.e., call inner pairs "outer pairs" and vice-versa. Neither the density ρ nor $|\Delta|$ will change. But both the sign of Δ and the value of β will be altered. With this in mind, it becomes clear that Eq. (3.35b) corresponds to the permutation symmetry, and that Eq. (3.35c) corresponds to the graph complement symmetry, both up to a renaming of the types.

[5] Another explanation is that there are effectively $q^* = 2$ pairs of blocks in the eyes of our formalism : A single inner pair and a single outer pair, with, respectively, a fraction β and $1-\beta$ of all possible edges.

3.9.5 Where the framework is put to the test : Inference

Procedure

It will be instructive to put our framework to the test and compare its predictions with numerical experiments that involve inference, i.e., the detection of the planted partition of actual instances of the GMGM. We will use the following procedure : (i) generate an instance of the model, (ii) run an inference algorithm on the instance, and (iii) compute the correlation of the inferred and planted partition (see below for a precise definition). The average detectability $\langle\mathcal{L}\rangle$ should bound the point where the average correlation becomes significant, and η–detectability should give an upper bound on the fraction of correlated instances.

Even for the small network size considered, it is impossible to compute all quantities involved in the process exactly ; we therefore resort to sub-sampling. We use an efficient algorithm[6] based on the Metropolis-Hastings algorithm of Ref. [224], which, unlike belief propagation [56], works well for dense networks with many short loops. The principle of the algorithm is to construct an ergodic chain of partitions $\mathcal{B}_0,...,\mathcal{B}_T$, and to sample from the chain to approximate the probability

$$\mu_i^r(G) = \sum_{\{\mathcal{B}_\sigma\}} \Pr(\mathcal{B}_\sigma|G,\boldsymbol{P},\boldsymbol{n})\delta(\sigma(v_i)=r) \tag{3.36}$$

that node v_i is in block B_r, given a network G and some parameter $\boldsymbol{P}$ and $\boldsymbol{n}$. It is easy to see that one can then maximize the probability of guessing the partition correctly by assigning nodes according to [55]

$$\hat{\sigma}(v_i) = \mathrm{argmax}_r(\mu_i^r)\ . \tag{3.37}$$

We choose a simple set of moves that yields an ergodic chain over all $\{\mathcal{B}\}$: at each step, we change the block of a randomly selected node v_i from $\sigma(v_i) = B_r$ to a randomly and uniformly selected block B_s, with probability $\min\{1,\mathcal{A}\}$, where

$$\mathcal{A} = \left[\frac{p_{rs}(1-p_{rr})}{p_{rr}(1-p_{rs})}\right]^{k_r^{(i)}}\left[\frac{p_{ss}(1-p_{rs})}{p_{rs}(1-p_{ss})}\right]^{k_s^{(i)}}\left[\frac{1-p_{rs}}{1-p_{rr}}\right]^{n_r-1}\left[\frac{1-p_{ss}}{1-p_{rs}}\right]^{n_s}\prod_{l\neq r,s}\left[\frac{p_{ls}(1-p_{rl})}{p_{rl}(1-p_{ls})}\right]^{k_l^{(i)}}\left[\frac{1-p_{ls}}{1-p_{rl}}\right]^{n_l} \tag{3.38}$$

and $k_l^{(i)}$ the number of neighbors of node v_i in block B_l [224]. The space of all partitions is obviously connected by this move set, and the possibility of resampling a configuration ensures that the chain is aperiodic. Furthermore, since transition probabilities are constructed according to the prescription Metropolis-Hastings, the chain is ergodic and samples from $\mathbb{P}(\mathcal{B}|G,\boldsymbol{P},\boldsymbol{n})$. Note that we assume that $\boldsymbol{P}$ is known when we compute Eq. (3.36). Learning the parameters can be done separately, see Ref. [55], for example.

In the spirit of Refs. [55, 56], we initialize the algorithm with the planted partition itself. This ensures that we will achieve the information-theoretic threshold, even if efficient in-

[6] We give a reference implementation of the algorithm in C++ at `www.github.com/jg-you/sbm_canonical_mcmc`.

ference is impossible [55]. To see this, first consider the case where the planted partition is information-theoretically detectable. In this case, the chain will concentrate around the initial configuration, and the marginal distribution [Eq. (3.36)] will yield a distribution correlated with the planted partition. We will have to proceed with care, however, since two scenarios may occur in the information-theoretically undetectable phase. If there is no hard phase—e.g., when $q = 2$ [159]—the algorithm will show no particular preference for the initial configuration and wander away toward partitions uncorrelated with the planted partition. But if there is a hard phase, one will have to wait for a period that diverges exponentially in the system size before the sampler becomes uncorrelated with its initial state [55]. This complicates convergence diagnosis and can lead one to conclude that correlated inference is possible even though it's not. To avoid these difficulties, we will simply restrict ourselves to the cases where the hard phase does not exist [1].

Once the estimated partition $\hat{\mathcal{B}}$ is obtained via Eq. (3.37), we compute its correlation with $\mathcal{B}$—the planted partition—using a measure that accounts for finite-size effects. The so-called relative normalized mutual information (rNMI) of Ref. [253] appears a good choice. Much like the well-known NMI [53, 132], the rNMI is bounded to the [0, 1] interval, and $\text{rNMI}(\mathcal{B}_p, \hat{\mathcal{B}}) = 1$ means that the planted partition $\mathcal{B}_p$ and the inferred partition $\hat{\mathcal{B}}$ are identical. Unlike the NMI, $\text{rNMI}(\mathcal{B}_p, \hat{\mathcal{B}}) = 0$ signals the absence of correlation between the two partitions, even in finite networks.

Results

In Fig. 3.4 (a), we plot $\langle \text{rNMI}(\mathcal{B}_p, \hat{\mathcal{B}}) \rangle$ in the density space of the GMGM. We use the parameters $\boldsymbol{W} = \boldsymbol{I}$, and $\boldsymbol{n} = [n/2, n/2]$ (i.e., the SSBM), since the resulting ensemble is conjectured to be the hardest of all, with respect to detectability [55]. Two important parallels can be drawn between the results shown in Fig. 3.4 (a) and the functional form of $\langle \mathcal{L}(\rho, \Delta; \beta) \rangle$ and $\eta(\rho, \Delta; \beta, n)$ [shown in Figs. 3.3 (b) and 3.3 (c) for a different GMGM]. First, notice how the boundary that marks the onset of the (theoretically) 1–detectable region partitions the density space in two qualitative regimes : A regime where perfect detection is possible *for all instances*, and a region where it is not. There is, of course, some level of arbitrariness involved in selecting the threshold T [see Eq. (3.28)]. But the fact that a line of constant η partitions the space is a hint that while $\mathcal{L} < 0$ is not sufficient for undetectability, there exists a level of significant λ^* for which $\mathcal{L}$ properly separates detectable and undetectable instances.

The second important parallel concerns hypersurfaces of constant $\langle \mathcal{L} \rangle$ and their connection with $\langle \text{rNMI} \rangle$. We have argued in Sec. 3.7 that $\langle \mathcal{L} \rangle$ is a good predictor of the accuracy of an optimal inference algorithms (with potentially exponential complexity). It should, therefore, not be surprising that there is an hypersurface of constant $\langle \mathcal{L} \rangle$ which *also* partitions the density space in two qualitative regions[7] : One where $\langle \text{rNMI} \rangle \approx 0$ and one where $\langle \text{rNMI} \rangle$

[7] We do not have a procedure to determine the value of λ within the information-theoretical framework

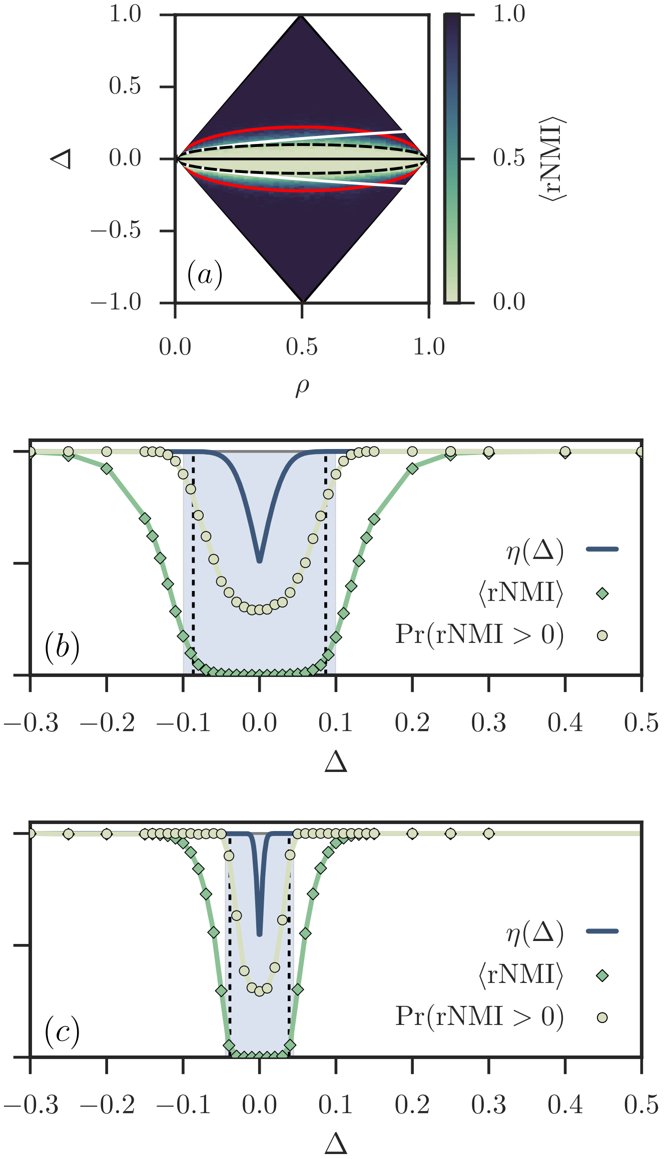


FIGURE 3.4 – Inference of the GMGM. All figures show results for the special case $q = 2$, $\boldsymbol{W} = \boldsymbol{I}_2$, and $\boldsymbol{n} = [n/2, n/2]$, corresponding to the $q = 2$ SSBM [1]. All empirical results are averaged over 10^4 independent instances of the SBM. (a) Average rNMI of the planted and the inferred partition in the density space of the model of size $n = 100$. Solid red lines mark the boundaries of the 1–detectability region, with tolerance threshold $T = 4\sqrt{2}$; see Eq. (3.28). Dotted black lines show the two solutions of $\Delta^*(\rho; \lambda = 1/2n, \beta)$; see Eq. (3.34). White lines show the finite-size Kesten-Stigum (KS) bound, before it is adjusted for the symmetries of the problem. (b, c) Phase transition at constant $\rho = 0.25$ for networks of $n = 100$ nodes (b) and $n = 500$ nodes (c). Circles indicate the fraction of instances for which a correlated partition could be identified, while diamonds show the average of the rNMI (lines are added to guide the eye). Blue solid curves show $\eta(\Delta; \rho, \beta, n)$; see Eq. (3.25). The shaded region lies below the finite-size KS bound $\Delta = \pm q\sqrt{\rho/n}$ (here with $q = 2$). The dotted lines show the two solutions of $\Delta^*(\rho; \lambda = 1/2n, \beta = 1/2)$.

is clearly greater than zero. On this hypersurface, the average level of significance is the same for all parametrizations of the GMGM; our results show that the inference algorithm achieves correspondingly uniform accuracy for all parameters on the surface.

One could argue that these parallels are not so obvious in Fig. 3.4 (a); we therefore focus on a subset of the density space in Figs. 3.4 (b) and 3.4 (c) to make our case clearer. In these figures, we plot the same information, but only for networks of constant density $\rho = 0.25$ and size $n = 100$ (b) and $n = 500$ (c). We also show the probability $\Pr(\mathrm{rNMI}(\mathcal{B}_p, \hat{\mathcal{B}}) > 0)$ that the inferred partition is correlated with the planted partition. This a direct measurement of the fraction of detectable instances, which we compare against $\eta(\Delta;\rho,\beta,n)$. It never reaches 0, because random fluctuations produce correlated partitions even when $\mathbb{P} = \mathbb{Q}$ (the rNMI corrects for the *average* correlation). If $\mathcal{L} > 0$ were a necessary and sufficient condition for detectability, then $\eta(\Delta;\rho,\beta,n)$ and $\Pr(\mathrm{rNMI} > 0|\Delta,\rho,\beta,n)$ would correspond perfectly. But since $\mathcal{L} > 0$ is only a *necessary* condition, $\eta(\Delta)$ acts as an upper bound rather than an exact expression, i.e., $\Pr(\mathrm{rNMI} > 0;\eta)$ can never be greater than $\eta(\Delta)$.

Two further observations must be made. First, it is known that in the sparse two-blocks SSBM, the transition between the information-theoretically undetectable and detectable regions occurs on the so-called Kesten-Stigum (KS) bound—located at $\Delta = \pm q\sqrt{\rho/n}$ for finite-size instances (this is not generally true, but the equivalence holds when $q = 2$ [159]). Despite the fact that this bound was derived for infinite ensembles, it holds very well in the finite case, as shown in Figs. 3.4 (b) and 3.4 (c). But the finite-size approach has the potential to be more precise. Building upon the interpretation of $\langle\mathcal{L}\rangle$ as a measure of the average difficulty of the inference problem, we set a threshold $\langle\mathcal{L}\rangle = 1/2n$ on the average detectability. For this choice of threshold, the approximate hypersurface equation predicts a transition at

$$\Delta^* = \pm 2\sqrt{\rho(1-\rho)/n}\,,$$

very close to the KS bound, but with a correction for nonvanishing densities. Interestingly, one can motivate this choice of threshold with random matrix theory [162, 191, 245] (see Appendix 3.11.2 for details) or the theory of low-rank matrix estimation [136]. The uncorrected and corrected bounds are shown on Fig. 3.4 (a). The corrected bound is qualitatively accurate in all density regimes, unlike the KS bound.

Second, in asymptotic theories, the SBM is either said to be undetectable with overwhelming probability, or the converse. The finite-size approach is more nuanced in the sense that it accounts for random fluctuations, which are also manifest in empirical results [see the curves $\Pr(\mathrm{rNMI}(\mathcal{B}_p, \hat{\mathcal{B}}) > 0)$]. While η–detectability is not perfect, as is argued above, it nonetheless goes through a smooth transition instead of an abrupt one. This reflects the continuous nature of the finite-size transition.

itself. However, random matrix theory and recent developments in Information theory offers some insights as to why one should have $\lambda \propto 1/n$, see Appendix 3.11.2 and Ref. [136] for details.

3.10 Conclusion

Building upon ideas from statistical theory, we have developed a framework to study the information-theoretic detectability threshold of the finite-size SBM. Our analysis relies on two different interpretations of the log-likelihood ratio $\mathcal{L}$ of the SBM and its equivalent random ensemble. We have used the rigorous interpretation of $\mathcal{L}$ to put a necessary condition on detectability. We have then computed the distribution of $\mathcal{L}$, and proved that up to half of the instances of the finite-size SBM could be declared undetectable on the basis of this simple test alone. We have further argued that the average of $\mathcal{L}$ could be interpreted as a proxy for the performance of an optimal inference algorithm (with possibly exponential running time). This interpretation has proved to be fruitful ; starting with a compact form for $\langle\mathcal{L}\rangle$, we have established the existence of a large equivalence class with respect to average detectability. In Appendix 3.11.1, we have shown that $\mathcal{L}$ can also be used to prove that, quite naturally, detectability decreases when the datasets are noisy. Using a correspondence with the finite-size information-theoretic threshold (as well as with random matrix theory, see Appendix 3.11.2), we have presented numerical evidence that the hypersurface $\langle\mathcal{L}\rangle = 1/2n$ separates detectable from undetectable instances in a special case of the SBM.

The unifying theme of this contribution has been the idea that $\langle\mathcal{L}\rangle$ quantifies both detectability and consistency in the finite-size SBM. This interpretation leaves many questions open for future works. Perhaps the most important of all : Can one determine the threshold within the framework of the theory itself, for general SBMs ?

A second important question pertains to sufficiency : Can one modify the condition to make it necessary *and* sufficient ? Or is a completely different approach needed ? In asymptotic analyses of the limit, one can use different conditions to bound the limit from above and below, as is done in Ref. [14]. Can a similar approach be fruitfully applied to finite instances ?

In closing, let us mention a few of the many possible generalizations of the methods introduced. First, it will be important to verify how our approach behaves in the limit $n \to \infty$. How this limit is taken will matter. In particular, we believe that our framework has much to say about the limit where $q \to \infty$, since it does not assume Poisson distributed degree, unlike other asymptotic theories of the limit. Second, we see no major obstacle to a generalization of our methods to other generative models of networks with a mesoscopic structure. This includes, for example, the consistency of graphons, a subject whose study has been recently undertaken [59]. Changing the null model from the equivalent random network ensemble to the configuration model [156, 178] could even allow an extension to degree-corrected SBM [116].

3.11 Appendix

3.11.1 Detectability and noise

One almost never has a perfect knowledge of the structure of real networks. The culprit can lie at the level of data collection, storage, transmission—or a combination of the above—, but the outcome is the same : Some edges are spurious and others are omitted [41]. To model imperfect knowledge, we will suppose that instances of the SBM first go through a noisy channel where T modifications—random edge removals or additions—are applied to the structure. Only then are we asked to tell which of hypotheses $\mathbb{P}$ and $\mathbb{Q}$ is the most likely. It should be clear that it will be more difficult to separate the two hypotheses, since noise is not necessarily aligned with the planted partition.

We will approach the problem with the following *universal perturbation process* (UPP) : At each step t of this process, a new random edge is added with probability c; otherwise, a random edge is removed. If a new edge must be added, then it is selected uniformly from the set of nonedges. If an edge must be removed, then it is selected uniformly from the set of edges already present in the network. This randomization step is then repeated T times. We call this process universal because one can map arbitrary perturbation patterns onto one or successive UPPs with different parameters c.

To prove that $\langle\mathcal{L}\rangle$ decreases as a result of any sufficiently long UPP, we will show that the total derivative

$$\frac{d}{dt}\langle\mathcal{L}\rangle = \sum_{r\leq s}\frac{\partial\langle\mathcal{L}\rangle}{\partial p_{rs}}\frac{dp_{rs}(t)}{dt} \tag{3.39}$$

is negative everywhere. In so doing, we assume that the process can be approximated as a continuous one (both with regards to “time” t and discrete quantities such as m_{rs}). Admittedly, a more rigorous approach would be needed to settle the matter unequivocally, but we argue that the method presented in this appendix gives a good intuition for the problem.

Without specifying the dynamics, and using Eq. (3.13), one can compute

$$\frac{\partial\langle\mathcal{L}\rangle}{\partial p_{rs}} = \alpha_{rs}\log\left[\frac{p_{rs}}{\rho}\frac{1-\rho}{1-p_{rs}}\right] = \alpha_{rs}x_{rs}\ , \tag{3.40}$$

where x_{rs} is identical to Eq. (3.23b). This leaves the $\dot{p}_{rs}(t)$ terms, whose expressions are determined by the perturbation dynamics. For the UPP, the evolution of $\{m_{rs}(t)\}_{r\leq s}$ is determined by the set of differential equations

$$\dot{m}_{rs}(t) = -\frac{(1-c)[m_{rs}(t)]}{\sum_{r\leq s} m_{rs}(t)} + \frac{c\ [m_{rs}^{\max} - m_{rs}(t)]}{m^{\max} - \sum_{r\leq s} m_{rs}(t)}. \tag{3.41}$$

The first term accounts for edge removal events, which occur with probability $(1-c)$ and involve edges that connect nodes in blocks (B_r, B_s) with probability $m_{rs}/\sum m_{rs}(t)$. A similar argument leads to the second term, which accounts for edge creation events.

Equation (3.41) can be transformed into an equation for $\dot{p}_{rs}(t)$ by dividing through by $m_{rs}^{\max}$, and then using the definitions $p_{rs}(t) = m_{rs}(t)/m_{rs}^{\max}$ and $\rho(t) = \sum_{r\leq s} m_{rs}(t)/m^{\max}$. We find

$$\dot{p}_{rs}(t) = \binom{n}{2}^{-1} \left[c\,\frac{1-p_{rs}(t)}{1-\rho(t)} - (1-c)\,\frac{p_{rs}(t)}{\rho(t)} \right] , \tag{3.42}$$

which, upon substitution in Eq. (3.39), yields

$$\frac{d\langle\mathcal{L}\rangle}{dt} = \Theta \sum_{r\leq s} \alpha_{rs} \log\left[\frac{f(p_{rs})}{f(\rho)}\right] \left[\frac{f(c)f(\rho)}{f(p_{rs})} - 1\right] , \tag{3.43}$$

where $\Theta = [2(1-c)p_{rs}]/[\rho n(n-1)]$ is a nonnegative factor, and where we have defined $f(x) = x/(1-x)$. It turns out that the sum is not only globally negative but that each term is also individually negative ; i.e.,

$$-\log\left[\frac{f(\rho)}{f(p_{rs})}\right] \left[\frac{f(c)f(\rho)}{f(p_{rs})} - 1\right] \leq 0 \qquad \forall r \leq s. \tag{3.44}$$

This comes about because the sign of the logarithm always matches that of the bracket.

To prove this statement, we treat five different cases and use the following identities repeatedly :

$$\frac{f(x)}{f(y)} < 1 \qquad \Longrightarrow \qquad x < y\,, \tag{3.45}$$

$$\frac{f(c)f(\rho)}{f(p_{rs})} > 1 \qquad \Longrightarrow \qquad c > \frac{p_{rs}(1-\rho)}{\rho(1-p_{rs}) + p_{rs}(1-\rho)}. \tag{3.46}$$

The cases are :

1. If $\rho = p_{rs}$: The logarithm equals 0 and the upper bound of Eq. (3.44) holds.
2. If $p_{rs} < \rho$ and $c < 1/2$: The logarithm is positive [see Eq. (3.45)]. The bracket is also positive, since the inequality in Eq. (3.46) can be rewritten as $(1-\rho)p_{rs} \leq \rho(1-p_{rs})$ using the fact that $c < 1/2$. This simplifies to $p_{rs} \leq \rho$, in line with our premise.
3. If $p_{rs} < \rho$ and $c \geq 1/2$: The logarithm is positive. Using our premise, we conclude that $f(\rho)/f(p_{rs}) > 1$ and $f(c) \geq 1$. Therefore, $f(c)f(\rho)/f(p_{rs}) > 1$, i.e., the bracket is positive.
4. If $p_{rs} > \rho$ and $c \leq 1/2$: The logarithm is negative. Using our premise, we conclude that $f(\rho)/f(p_{rs}) < 1$ and $f(c) \leq 1$. Therefore, $f(c)f(\rho)/f(p_{rs}) < 1$, i.e., the bracket is negative.
5. If $p_{rs} > \rho$ and $c > 1/2$: The logarithm is negative. The bracket is also negative, since the converse of the inequality in Eq. (3.46) can be rewritten as $(1-\rho)p_{rs} \geq \rho(1-p_{rs})$ using the fact that $c > 1/2$. This simplifies to $p_{rs} \geq \rho$, in line with our premise.

This list covers all cases and therefore completes the proof that $d\langle\mathcal{L}\rangle/dt \leq 0$, i.e., that average detectability decreases as a result of the application of a UPP.

3.11.2 Connection with random matrix theory

In Refs. [162, 191] it is argued that SBM is not efficiently detectable when the extremal eigenvalues of the modularity matrix of its instances merge with the so-called "continuous eigenvalue band." It is proved in Ref. [162] that this occurs when

$$n(p_{\text{in}} - p_{\text{out}}) = \pm\frac{1}{n}\sqrt{2n(p_{\text{in}} + p_{\text{out}})}\,, \tag{3.47}$$

for the two-block SSBM with Poisson distributed degrees. Furthermore, in this case, there is no so-called hard phase [159], meaning that the above limit affords a comparison with the prediction if our information theoretic framework.

Since we are concerned with the finite case, let us first modify this result to account for binomial distributed degrees instead. It turns out that the corrected condition is found by substituting the expectations of Poisson variables [in the right-hand-side of Eq. (3.47)] by that of binomial variables. This leads to

$$(p_{\text{in}} - p_{\text{out}}) = \pm\frac{1}{n}\sqrt{2n[p_{\text{in}}(1-p_{\text{in}}) + p_{\text{out}}(1-p_{\text{out}})]}\,, \tag{3.48}$$

or, in terms of the natural parameters of the GMGM,

$$\Delta^* = \pm\sqrt{\frac{4}{n-1}\rho(1-\rho)}\,. \tag{3.49}$$

This equation bears a striking similarity with Eq. (3.34), our approximate equation for curves of constant $\langle\mathcal{L}\rangle$. In fact, for the two-block SSBM ($\beta \approx 1/2$), the latter reads

$$\Delta^* = \pm\sqrt{8\lambda\rho(1-\rho)}\,. \tag{3.50}$$

One obtains an exact equivalence between the two expressions by setting $\lambda = 1/2(n-1) \approx 1/2n$. The fact that modularity based spectral methods cannot infer a correlated partition if $\Delta \leq \Delta^*$ [Eq. (3.49)] can thus be understood as stemming from a lack of statistical evidence for the SBM.

3.11.3 Detailed proofs : Symmetries of the average detectability

Theorem 1 (λ–preserving symmetries)**.** *All transformations $T(\boldsymbol{\alpha}, \boldsymbol{P})$ of the parameter space of the SBM that are (i) reversible, (ii) space-preserving, and (iii) valid at every point of the parameter space can be written as*

$$p_{rs} \mapsto p'_{rs} = \gamma_{rs} + (1 - 2\gamma_{rs})p_{\omega(r,s)}\,, \tag{3.51a}$$

$$\alpha_{rs} \mapsto \alpha'_{rs} = \alpha_{\pi(r,s)}\,, \tag{3.51b}$$

where $\gamma_{rs} \in \{0,1\}$ and where π and ω are permutations that acts on the set $\{(r,s)\,|\,1 \leq r, \leq s \leq g\}$. Under the additional constraint that $\langle\mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{P})\rangle$ be preserved by $\{T\}$ and equal to λ, one must have

$$\pi = \omega \quad \textit{and} \quad \gamma_{rs} = \gamma \quad \forall (r,s)\,.$$

Let us first introduce new notations to clarify the proof of Theorem 1. First, we define vectors $|p\rangle$ and $|\alpha\rangle$ whose entries are the $q^* = \binom{q}{2} + q$ entries of the upper triangle (and diagonal) of $\boldsymbol{P}$ and $\boldsymbol{\alpha}$. In this notation, we write the average density as $\langle\alpha|p\rangle$ and the average detectability as

$$\langle\mathcal{L}(\boldsymbol{\alpha},\boldsymbol{P})\rangle = \langle\alpha|u(\boldsymbol{\alpha},\boldsymbol{P})\rangle \ , \tag{3.52}$$

where $|u(\boldsymbol{\alpha},\boldsymbol{P})\rangle$ is q^*–dimensional vector parametrized by $(\boldsymbol{\alpha},\boldsymbol{P})$, whose entries are given by

$$u_{rs}(\boldsymbol{\alpha},\boldsymbol{P}) = p_{rs}\log\frac{p_{rs}}{\langle\alpha|p\rangle} + (1-p_{rs})\log\frac{1-p_{rs}}{1-\langle\alpha|p\rangle} \ .$$

We also introduce $\boldsymbol{\Pi}$ and $\boldsymbol{\Omega}$, two $q^* \times q^*$ permutation matrices such that $\boldsymbol{\Pi}\,|\alpha\rangle_{rs} = \alpha_{\pi(r,s)}$ and $\boldsymbol{\Omega}\,|p\rangle_{rs} = p_{\omega(r,s)}$, where $|a\rangle_{ij}$ is the element (i,j) of vector $|a\rangle$. In this notation, Eqs. (3.51) are given by

$$\begin{aligned}
|\alpha\rangle &\mapsto |\alpha'\rangle = \boldsymbol{\Pi}\,|\alpha\rangle \ , \\
|p\rangle &\mapsto |p'\rangle = \boldsymbol{\Gamma}\,|1\rangle + (\boldsymbol{I} - 2\boldsymbol{\Gamma})\boldsymbol{\Omega}\,|p\rangle \\
&\qquad\quad \equiv \boldsymbol{\Omega}\boldsymbol{\Gamma}'\,|1\rangle + \boldsymbol{\Omega}(\boldsymbol{I} - 2\boldsymbol{\Gamma}')\,|p\rangle \ ,
\end{aligned}$$

where $\boldsymbol{\Gamma}$ is a diagonal matrix with element γ_{rs} on the diagonal, where $\boldsymbol{I}$ is the identity matrix, and where $\boldsymbol{\Gamma}' = \boldsymbol{\Omega}^{-1}\Gamma$ is also a diagonal matrix.

Proof. The proof of the first part of Theorem 1 (form of the transformations) is given in the main text, see Sec. 3.7.3.

To prove the second part of the theorem (constrained transformations), we look for the subset of all transformations of the form shown in Eq. (3.51) that also preserve $\langle\mathcal{L}\rangle$, i.e., transformations T in $S_{q^*} \times B_{q^*}$ that map $(\boldsymbol{\alpha},\boldsymbol{P})$ to $(\boldsymbol{\alpha}',\boldsymbol{P}')$ and that satisfy

$$\langle\alpha|u(\boldsymbol{\alpha},\boldsymbol{P})\rangle = \langle\alpha'|u(\boldsymbol{\alpha}',\boldsymbol{P}')\rangle \ .$$

It is easy to check that if $\boldsymbol{\Omega} = \boldsymbol{\Pi}$ and $\boldsymbol{\Gamma} = \gamma\boldsymbol{I}$ with $\gamma \in \{0,1\}$, then the average density and the normalized log-likelihood are both preserved. Therefore, if the transformations are of the proposed form, then λ is preserved.

To complete the proof we must show that $\langle\mathcal{L}\rangle$ is conserved *only if* $\boldsymbol{\Gamma} = \gamma\boldsymbol{I}$ and $\boldsymbol{\Omega} = \boldsymbol{\Pi}$. First, we note that by the properties of the scalar product and permutation matrices, we have the following obvious symmetry

$$\langle\alpha|u\rangle = \langle\boldsymbol{\Pi}\alpha|\boldsymbol{\Pi}u\rangle \, ,$$

which is valid for all permutation matrices $\boldsymbol{\Pi}$. We use this symmetry to "shift" all permutation matrices to the second part of the scalar product representation of $\langle\mathcal{L}\rangle$, i.e., we write

$$\langle\alpha|u\rangle \mapsto \langle\alpha'|u'\rangle = \langle\boldsymbol{\Pi}\alpha|u'\rangle = \langle\alpha|\boldsymbol{\Pi}^{-1}u'\rangle \ .$$

Now, from Eq. (3.52), it is clear that we will have $\langle \mathcal{L}(\boldsymbol{\alpha},\boldsymbol{P})\rangle = \langle \mathcal{L}(\boldsymbol{\alpha}',\boldsymbol{P}')\rangle$ if and only if

$$\langle \alpha | u - \boldsymbol{\Pi}^{-1} u' \rangle = 0 \,, \tag{3.53}$$

where $|u'\rangle := |u(\boldsymbol{\alpha}',\boldsymbol{P}')\rangle$. Since $|u - \boldsymbol{\Pi}^{-1}u'\rangle$ is analytic in $\boldsymbol{\alpha}$, we can expand it by using Taylor series ; this creates an infinite series of constraints that must all be satisfied. In particular, the condition in Eq. (3.53) will be satisfied only if

$$|u - \boldsymbol{\Pi}^{-1}u'\rangle = |0\rangle \ .$$

This is true if and only if, for all (r,s), one has

$$p_{rs} \log \frac{p_{rs}}{\langle \alpha | p \rangle} + (1-p_{rs}) \log \frac{1-p_{rs}}{1-\langle \alpha | p \rangle} = \bar{p}_{rs} \log \frac{\bar{p}_{rs}}{\langle \alpha | \bar{p} \rangle} + (1-\bar{p}_{rs}) \log \frac{1-\bar{p}_{rs}}{1-\langle \alpha | \bar{p} \rangle} \,, \tag{3.54}$$

where $|\bar{p}\rangle = \boldsymbol{\Pi}^{-1}|p'\rangle$. Here, $|\bar{p}\rangle$ is the transformed vector $|p'\rangle$, on which the inverse of permutation $\pi(r,s)$ is also applied.

Let us now suppose that $\boldsymbol{\alpha}$ tends to the point $\tilde{\boldsymbol{\alpha}}$, which is such that $\tilde{\alpha}_{rs} = 0$ for all (r,s) except for $(r,s) = (a,b)$ (i.e., $\tilde{\alpha}_{ab} = 1$). In this limit, Eq. (3.54) is trivially satisfied when $(r,s) = (a,b)$ but not otherwise. Let us suppose $(r,s) \neq (a,b)$ and expand the equation around $p_{ab} = \bar{p}_{ab} = \frac{1}{2}$. From this second series expansion, one concludes that the equality is satisfied if either $\bar{p}_{ab} = p_{ab}$ or $\bar{p}_{ab} = 1 - p_{ab}$. In both cases, the indices must match, which implies that $(a,b) = \pi^{-1} \circ \omega(a,b)$. By repeating the same argument for all (a,b), we conclude that $\omega = \pi$. Thus, the map $T : (\boldsymbol{\alpha},\boldsymbol{P}) \mapsto (\boldsymbol{\alpha}',\boldsymbol{P}')$ is a symmetry only if $\boldsymbol{\Pi} = \boldsymbol{\Omega}$.

This leaves the proof that $\boldsymbol{\Gamma} = \gamma \boldsymbol{I}$. Let us, by contradiction, assume that γ_{rs} differs from one set of indices to the other and define the sets A and B by

$$A = \{(r,s) : \gamma_{rs} = 0\} \quad \text{and} \quad B = \{(r,s) : \gamma_{rs} = 1\} \,.$$

Then one can write

$$\rho = \langle \alpha | p \rangle = \langle p \rangle_A + \langle p \rangle_B \,, \tag{3.55}$$

where $\langle p \rangle_X := \sum_{(r,s)\in X} \alpha_{rs} p_{rs}$. Returning to Eq. (3.54) for $(r,s) \in A$ and using the newfound fact that $\boldsymbol{\Pi} = \boldsymbol{\Omega}$ which implies $\bar{p}_{rs} = \gamma_{rs} + (1-2\gamma_{rs})p_{rs}$ (no more permutations), we find

$$p_{rs} \log \frac{p_{rs}}{\rho} + (1-p_{rs}) \log \frac{1-p_{rs}}{1-\rho} = p_{rs} \log \frac{p_{rs}}{\langle p' \rangle_A + \langle p' \rangle_B} + (1-p_{rs}) \log \frac{1-p_{rs}}{1-\langle p' \rangle_A - \langle p' \rangle_B} \,.$$

This can only be true if $\rho = \langle p' \rangle_A + \langle p' \rangle_B$, i.e., if $A = \varnothing$ or $B = \varnothing$. Therefore, $\gamma_{rs} = \gamma \ \forall (r,s)$, with $\gamma \in \{0,1\}$.

□

3.11.4 Detailed proofs : Convexity of $\langle\mathcal{L}\rangle$

Theorem 2. $\langle\mathcal{L}(\boldsymbol{\alpha},\boldsymbol{P})\rangle$ *is convex with respect to* $\boldsymbol{P}$*.*

This property of $\langle\mathcal{L}\rangle$ is—perhaps surprisingly—not a consequence of the convexity of the KL divergence. Instead, it follows from the log-sum inequality.

Proof. We prove that $\langle\mathcal{L}(\boldsymbol{\alpha},\boldsymbol{P})\rangle$ is convex with respect to $\boldsymbol{P}$ by showing that it satisfies the convexity condition

$$\langle\mathcal{L}(\boldsymbol{\alpha},(1-t)\boldsymbol{P}+t\boldsymbol{Q})\rangle \leq (1-t)\langle\mathcal{L}(\boldsymbol{\alpha},\boldsymbol{P})\rangle + t\langle\mathcal{L}(\boldsymbol{\alpha},\boldsymbol{Q})\rangle\ , \tag{3.56}$$

explicitly for all $t\in[0,1]$. Again, for the sake of clarity, we will use the notation developed in the previous section, and, in particular, write the density as $\rho=\langle\alpha|p\rangle$. We write each term on the left-hand-side of Eq. (3.56) as

$$\begin{aligned}\alpha_{rs}\Bigg\{&[(1-t)p_{rs}+tq_{rs}]\log\frac{(1-t)p_{rs}+tq_{rs}}{(1-t)\,\langle\alpha|p\rangle+t\,\langle\alpha|q\rangle}\\&+[(1-t)(1-p_{rs})+t(1-q_{rs})]\log\frac{(1-t)(1-p_{rs})+t(1-q_{rs})}{(1-t)(1-\langle\alpha|p\rangle)+t(1-\langle\alpha|q\rangle)}\Bigg\}\end{aligned} \tag{3.57}$$

It is easy to see that the log-sum inequality

$$(a+\bar{a})\log\frac{a+\bar{a}}{b+\bar{b}}\leq a\log\frac{a}{b}+\bar{a}\log\frac{\bar{a}}{\bar{b}}$$

can be applied to both parts of Eq. (3.57) to separate terms by their coefficients $(1-t)$ and t. Repeating the same operation on all terms yields the inequality in Eq. (3.56). □

3.12 Supplementary Material

3.12.1 From the KL divergence to difference of entropies

In this short section, we give an alternative proof of the compact form $\langle\mathcal{L}\rangle \propto D(\mathbb{P}||\mathbb{Q}) = m^{\max}[h(\rho) - \sum_{r\leq s}\alpha_{rs}h(p_{rs})]$, based on simple argument from information-theory.

Proof. The following chain rule holds for the Kullback-Leibler divergence [48] :

$$D\big(p(x,y)||q(x,y)\big) = D\big(p(x)||q(x)\big) + D\big(p(y|x)||q(y|x)\big)\ ,$$

where p,q are joint distributions for (X,Y) on the same supports $\mathcal{X},\mathcal{Y}$. It is easy to see that if X and Y are independent according to both distributions p and q, then the divergence is additive, i.e,

$$D\big(p(x,y)||q(x,y)\big) = D\big(p(x)||q(x)\big) + D\big(p(y)||q(y)\big)\ .$$

This property trivially generalizes to an arbitrary number of random variables [48]. Because the distribution over all graphs can be seen as a product of independent distributions over all edges, we may write the Kullback-Leibler divergence as

$$D\big(\mathbb{P}||\mathbb{Q}\big) = \sum_{i<j} D\big(P_{ij}||Q_{ij}\big)\ , \tag{3.58}$$

where P_{ij} and Q_{ij} are the Bernoulli random variables that govern the existence of the edge linking nodes v_i and v_j, according to the distribution $\mathbb{P}$ and $\mathbb{Q}$. These variables have two outcomes : Either the edge exists with probability $p_{\sigma_i\sigma_j}$ (resp. $q_{\sigma_i\sigma_j}$), or does not with probability $1-p_{\sigma_i\sigma_j}$ (resp. $1-q_{\sigma_i\sigma_j}$). We write σ_i as an abbreviation of $\sigma(v_i)$, the index of the block to which node v_i belongs. The divergence associated with an edge is therefore

$$\begin{aligned} D\big(P_{ij}||Q_{ij}\big) &= p_{\sigma_i\sigma_j}\log\frac{p_{\sigma_i\sigma_j}}{q_{\sigma_i\sigma_j}} + (1-p_{\sigma_i\sigma_j})\log\frac{1-p_{\sigma_i\sigma_j}}{1-q_{\sigma_i\sigma_j}} \\ &= -h(p_{\sigma_i\sigma_j}) - p_{\sigma_i\sigma_j}\log q_{\sigma_i\sigma_j} - (1-p_{\sigma_i\sigma_j})\log(1-q_{\sigma_i\sigma_j})\ , \end{aligned}$$

where $h(x)$ is, again, the binary entropy. Now, taking $q_{\sigma_i\sigma_j} = \rho$ for all pairs and summing over all edges [see Eq. (3.58)], we find, after grouping terms by types,

$$D\big(\mathbb{P}||\mathbb{Q}\big) = -\sum_{r\leq s} m_{rs}^{\max} h(p_{rs}) - \sum_{r\leq s} m_{rs}^{\max}[p_{rs}\log\rho + (1-p_{rs})\log(1-\rho)]\ .$$

The final result follows from the definition of the average density, after one substitutes $m_{rs}^{\max}$ for $\alpha_{rs}m^{\max}$. □

3.12.2 Another case study : The general two blocks SBM

In the main text, we have analyzed the general modular graph model of Ref. [121] and have shown how our framework yielded nicely interpretable expressions for such a low-dimensional model. We now investigate another simple model—the general two blocks SBM—which also lends itself to exhaustive analysis.

Parameterizing the two blocks SBM

The parametrization of the two blocks SBM is specified with the 3 probabilities $\{p_{11}, p_{22}, p_{12}\}$ as well as the two blocks sizes, n_0 and n_1. It contains the core-periphery model (with $p_{11} > p_{12}$ and $p_{22} \approx 0$) and the SSBM ($p_{11} = p_{22}$) as special cases. We follow the path laid out in our analysis of the GMGM and obtain a natural parametrization in terms of average density and displacement from the equivalent random model. The simplest parameterization of this kind is perhaps

$$\rho = \alpha_{11} p_{11} + \alpha_{22} p_{22} + \alpha_{12} p_{12} \,, \tag{3.59a}$$

$$\Delta_x = p_{22} - p_{11} \,, \tag{3.59b}$$

$$\Delta_y = -\frac{\alpha_{11}}{1-\alpha_{12}} p_{11} - \frac{\alpha_{22}}{1-\alpha_{12}} p_{22} + p_{12} \,, \tag{3.59c}$$

Figure 3.5 (a) illustrates the change of variables. It amounts to placing ourselves in a plane of constant density in the $p_{11} \times p_{22} \times p_{12}$ space, then measuring the distance from the point $(p_{11}, p_{22}, p_{12}) = (\rho, \rho, \rho)$ in the Δ_x and Δ_y direction. Again, much like in the case of the GMGM, not all $(\rho, \Delta_x, \Delta_y)$ tuples correspond to probabilities—it is only true for the region of the constant density plane contained inside $[0, 1]^3$. When ρ is close to 0 or 1, this region is a triangle. At intermediate values of ρ, parts of the triangle lie outside the cube of probabilities—the region is then a polygon of more than 3 edges, see Fig. 3.5 (b) for an example.

We do not investigate the general 2 blocks SBM as thoroughly as the GMGM, since much of the results and observations are identical to those of the previous case study. It will, however, be instructive to consider the average detectability and symmetries of this simple model.

Average detectability

The average log-likelihood $\langle \mathcal{L} \rangle$ is, in the $\rho \times \Delta_x \times \Delta_y$ space,

$$\begin{aligned}\langle \mathcal{L}(\rho, \Delta_x, \Delta_y, \beta) \rangle = \quad & \beta^2 \left\{ h(\rho) - h\left[\rho - \frac{(1-\beta)^2}{1-2\beta(1-\beta)}\Delta_x - 2\beta(1-\beta)\Delta_y\right]\right\} \\ & + (1-\beta)^2 \left\{ h(\rho) - h\left[\rho + \frac{\beta^2}{1-2\beta(1-\beta)}\Delta_x - 2\beta(1-\beta)\Delta_y\right]\right\} \\ & + 2\beta(1-\beta)\left\{h(\rho) - h\left[\rho + [1-2\beta(1-\beta)]\Delta_y\right]\right\} \,. \end{aligned} \tag{3.60}$$

Our main objective in studying the two-blocks SBM is to showcase anew the technique introduced in the main text to compute the hypersurfaces of constant $\langle \mathcal{L} \rangle$. The first step is to invert the parametrization (3.59)

$$p_{11} = \rho - \frac{\alpha_{22}}{1-\alpha_{12}} \Delta_x - \alpha_{12} \Delta_y \,, \tag{3.61}$$

$$p_{22} = \rho + \frac{\alpha_{11}}{1-\alpha_{12}} \Delta_x - \alpha_{12} \Delta_y \,, \tag{3.62}$$

$$p_{12} = \rho + (1-\alpha_{12}) \Delta_y \,. \tag{3.63}$$

Upon substitution of these parameters in

$$2\lambda\rho(1-\rho) = \sum_{r\leq s}\alpha_{rs}(p_{rs}-\rho)^2 + \mathcal{O}[(p_{rs}-\rho)^3]\ ,$$

one finds

$$\begin{aligned} 2\lambda\rho(1-\rho) &\approx \sum_{r\leq s}\alpha_{rs}(p_{rs}-\rho)^2\ ,\\ &= \Delta_x^2\left[\frac{\alpha_{11}\alpha_{22}}{(1-\alpha_{12})}\right] + \Delta_y^2[\alpha_{12}(1-\alpha_{12})]\ . \end{aligned} \tag{3.64}$$

Equation (3.64) predicts that these hypersurfaces are ellipses in the plane of constant density ρ, centered at $(\Delta_x = 0, \Delta_y = 0)$, with major axis $\alpha_{11}\alpha_{22}/(1-\alpha_{12})$ and minor axis $\alpha_{12}(1-\alpha_{12})$. This result is put to the test in Fig. 3.5 (b), where we also show numerical solutions for comparison. Agreement is, once again, excellent when λ is not too large, while $\lambda \gg 1$ leads to significant errors in our prediction. Following the ellipses in Fig. 3.5-(c) shows that the easiest inference problems—those where $\langle\mathcal{L}\rangle$ is the largest—are the ones in the corner on the edges of the accessible region. These regions are those where block pairs are well segregated, i.e., the bipartite ensemble with $(p_{11}, p_{22}, p_{12}) = (0, 0, \rho)$ [top-corner], the perfectly assortative cases with $(p_{11}, p_{22}, p_{12}) = (a\rho, (1-a)\rho, 0)$ where a is a constant [bottom edge].

Symmetries

The purpose of the parametrization $(\rho, \Delta_x, \Delta_y)$ is to facilitate the calculation of hypersurfaces of constant $\langle\mathcal{L}\rangle$. For the GMGM, the natural parametrization also had the added benefit of highlighting the symmetries naturally. This needs not be the case in all variants of the SBM, as we now show.

Let us first simplify the notation; we define β as the fraction of nodes that belong to block B_1, i.e., $n_1 = \beta n$ and $n_2 = (1-\beta)n$ (not to be confused with the β of the GMGM). In the limit where $n \gg 1$, we then have

$$\alpha_{11} \approx \beta^2 \quad \alpha_{22} \approx (1-\beta)^2 \quad \alpha_{12} \approx 2\beta(1-\beta)\ . \tag{3.65}$$

Since the parametrization (3.59) assigns a special significance to the (p_{11}, p_{22}) direction, the symmetries that involve these two blocks [and *not* the (B_1, B_2) pair], are the simplest. By direct enumeration, one finds

$$(\rho, \Delta_x, \Delta_y, \beta) \mapsto (\rho, \Delta_x, \Delta_y, \beta)\ , \tag{3.66a}$$

$$(\rho, \Delta_x, \Delta_y, \beta) \mapsto (1-\rho, -\Delta_x, -\Delta_y, \beta)\ , \tag{3.66b}$$

$$(\rho, \Delta_x, \Delta_y, \beta) \mapsto (\rho, -\Delta_x, \Delta_y, 1-\beta)\ , \tag{3.66c}$$

$$(\rho, \Delta_x, \Delta_y, \beta) \mapsto (1-\rho, \Delta_x, -\Delta_y, 1-\beta)\ , \tag{3.66d}$$

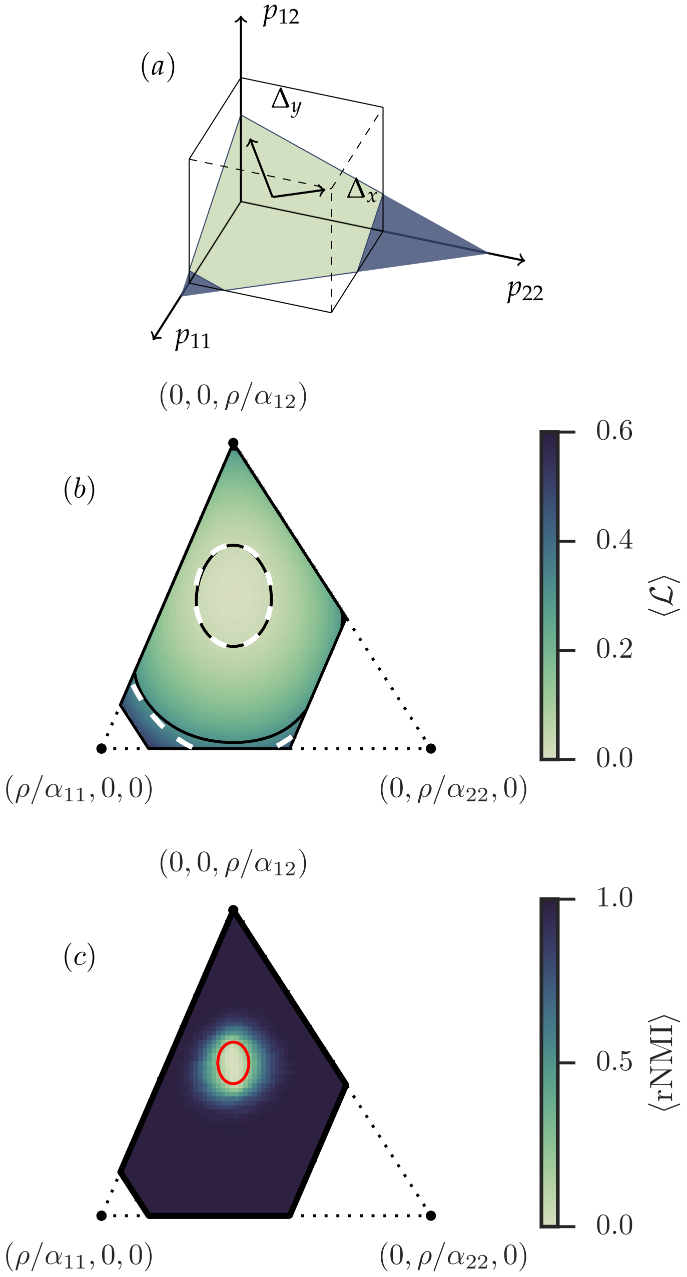


FIGURE 3.5 – Detectability of the general two-blocks SBM. We use the size vector $\boldsymbol{n} = [55, 45]$ and density $\rho = 0.35$ throughout the figure. (a) Natural parametrization of the two blocks SBM. A plane of constant density is shown; the lighter region of this plane is inside $[0,1]^3$. Axes Δ_x and Δ_y lie *on* the plane; their origin is the point $(p_{11}, p_{22}, p_{12}) = (\rho, \rho, \rho)$. (b) $\langle\mathcal{L}(\Delta_x, \Delta_y; \rho, \boldsymbol{\alpha})\rangle$ in the plane of constant density. Corners are identified by their coordinates in the $p_{11} \times p_{22} \times p_{12}$ space. Both the numerical solutions of $\langle\mathcal{L}\rangle = \lambda$ (solid black line) and the predictions of Eq. (3.64) (dashed line) are shown, for $\lambda = 0.12$ and 0.1. (c) Average rNMI of the planted and the inferred partition in the plane of constant density. The solid red line shows the solutions of $\langle\mathcal{L}\rangle = 1/2n$.

i.e., the identity, the pure graph complement, the permutation of block pairs (0,0) and (1,1), and the same permutation accompanied by the graph complement. Notice how this subset of transformation forms, once again, a group isomorphic to the Klein four-group.

The transformation equations would be, however, much less compact if we were to list transformations involving, say, cyclic permutations of the blocks. This situation is common : Choices of parameters that favor a particular pair of blocks will yield compact symmetry equations for this pair of blocks, but not the others. Therefore symmetries are, in general, best expressed directly in terms of $\boldsymbol{\alpha}$ and $\boldsymbol{P}$, unless the model has a special and all encompassing parametrization—like the GMGM.

Deuxième partie

Inférence temporelle

Chapitre 4

Inférence temporelle : Archéologie des réseaux

Article original :

“Network archaeology : phase transition in the recoverability of network history”

Jean-Gabriel Young [1, 2], Laurent Hébert-Dufresne [1, 3, 4], Edward Laurence [1, 2], Charles Murphy [1, 2], Guillaume St-Onge [1, 2] et Patrick Desrosiers [1, 2, 5]

[1] Département de Physique, de Génie Physique, et d’Optique
Université Laval, Québec (QC), G1V 0A6, Canada
[2] Centre interdisciplinaire de modélisation mathématique de l’Université Laval
Québec (QC), G1V 0A6, Canada
[3] Department of Computer Science, University of Vermont, Burlington, VT, 05405, USA
[4] Vermont Complex Systems Center, University of Vermont, Burlington, VT, 05405, USA
[5] Centre de recherche de CERVO, Québec (QC), G1J 2G3, Canada

Référence : arXiv:1803.09191 (2018) [249] (§4.3–4.8.7) [‡]

[‡] Ces sections sont reproduites directement de l’article orignal. Le contenu n’en a été modifié que pour se conformer au format exigé par la Faculté des études supérieures et postdoctorales de l’Université Laval.

4.1 Avant-propos

Dans les deux chapitres de la première partie, nous avons vu qu'il était possible d'extraire de l'information *structurelle* d'un réseau, en supposant qu'il possède une structure latente. La clé de voûte de ce travail numérique et mathématique était un modèle aléatoire (le SBM), qui nous a permis de relier les observations (le réseau) aux quantités recherchées (les blocs). Ce modèle a été un premier exemple de l'approche dite par modèles génératifs à l'inférence statistique.

Bien que historiquement limités à la détection de groupes et au plongement de réseaux dans des espaces simples [110], les modèles génératifs sont plus flexibles qu'on pourrait le penser. Une série de résultats très récents montrent que les modèles génératifs peuvent être utilisés pour, en autres : retracer la source d'une épidémie [143, 150], ou découvrir la hiérarchie naturelle des noeuds d'un réseau dirigé [54]. Encouragé par ces résultats, on se propose dans cette deuxième partie de développer une nouvelle application des modèles génératifs, soit l'inférence *temporelle*.

On supposera donc qu'un réseau est généré par un modèle de croissance (ici le modèle PA généralisé [128, 129]), de sorte qu'il évolue dans le temps. On reliera cette fois l'observable à des quantités cachées temporelles. On verra donc la structure d'un réseau statistique comme une signature de son passé. Notre objectif sera de reconstruire cette histoire.

Ce problème n'a été que peu étudié jusqu'à présent, probablement à cause de son apparente complexité. Supposant qu'un réseau évolue par ajout séquentiel de M liens, on peut énumérer jusqu'à $M!$ histoires différentes. L'espace des solutions est donc énorme—beaucoup plus grand que ceux étudiés dans les chapitres précédents, dont la taille était seulement exponentielle. Trouver l'histoire la plus vraisemblable semble donc être une tâche ardue. Or, comme le montrent les travaux du présent chapitre, le problème de reconstruction temporelle est en fait relativement simple.

4.2 Résumé

Les processus de croissance réseaux peuvent être vus comme des processus génératifs pour la structure *et* l'histoire des réseaux complexes. Ce point de vue mène naturellement au problème de l'archéologie réseau, soit la reconstruction de tous les états antérieurs d'un réseau en croissance, à partir de sa structure observée. Dans ce chapitre, on introduit une formulation bayésienne de l'archéologie réseau. On utilise une généralisation du modèle d'attachement préférentiel (PA) en guise de modèle génératif. On développe un algorithme d'échantillonnage séquentiel préférentiel pour évaluer la distribution a posteriori sur les histoires possibles d'un réseau, ainsi qu'une méthode heuristique efficace qui approxime les moyennes de cette distribution, en temps linéaire. Ces deux méthodes nous permettent d'identifier et de caractériser une transition de phase. Plus précisément, pour des réseaux construits à l'aide du modèle génératif lui-même, on montre qu'il existe une phase où la reconstruction de leur histoire est impossible, et une phase où cette reconstruction est possible. On conclut avec une étude de cas, où on démontre l'utilité de la méthode en extrayant partiellement l'histoire d'un réseau réel, qui n'est pas explicitement généré par le modèle PA.

4.3 Abstract

Network growth processes can be understood as generative models of the structure and history of complex networks. This point of view naturally leads to the problem of network archaeology : Reconstructing all the past states of a network from its structure—a difficult permutation inference problem. In this paper, we introduce a Bayesian formulation of network archaeology, with a generalization of preferential attachment as our generative mechanism. We develop a sequential importance sampling algorithm to evaluate the posterior averages of this model, as well as an efficient heuristic that uncovers the history of a network in linear time. We use these methods to identify and characterize a phase transition in the quality of the reconstructed history, when they are applied to artificial networks generated by the model itself. Despite the existence of a no-recovery phase, we find that non-trivial inference is possible in a large portion of the parameter space as well as on empirical data.

4.4 Introduction

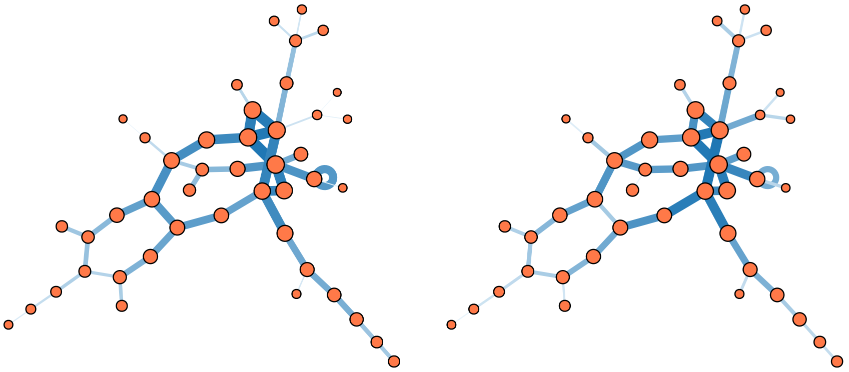
Unequal distributions of resources are ubiquitous in the natural and social world [100]. While inequalities abound in many contexts, their impact is particularly dramatic in complex networks, whose structure is heavily constrained in the presence of skewed distributions. For instance, the aggregation of edges around a few hubs determines the outcome of diseases spreading in a population [226], the robustness of technological systems to targeted attacks and random failures [35], or the spectral property of many networks [37]. It is therefore not surprising that much effort has been devoted to understanding how unfair distributions come about in networks. Many of the satisfactory explanations thus far uncovered have taken the form of constrained growth processes : the rich-get-richer principle [16], sampling space reduction processes [230] and latent fitness models [21] are all examples of growth processes that lead to a heavy-tailed distribution of the degrees.

A common characteristic shared by these processes is that they do not—nor are they expected to—give a perfect account of reality [222]. Their rules are simple, and only capture the essence of the mechanisms at play, glossing over details [155]. But despite these simplifications, growth processes endure as useful models of real complex systems. At a macroscopic level, their predictions have often been found to fit the statistics of real networks to surprising degrees of accuracy [102]. At a microscopic level, they have been shown to act effectively as *generative models* of complex networks [87, 137], i.e., as stochastic processes that can explain the details of a network's structure [110]. This point of view has led, for example, to powerful statistical tests that delineate the role of nodes' fitness and age in determining a network's structure [21, 201].

The notion of growth processes as generative model is now being pushed further than ever before. The burgeoning field of *network archaeology* [164], in particular, builds upon the idea that growth processes are generative models of the *history* of complex networks, able to reveal the past states of static systems. This point of view is perhaps the most clearly stated in the bioinformatics literature, which seeks to reconstruct ancient protein–protein interaction (PPI) networks to, e.g., improve PPI network alignment algorithms [66, 75] or understand how the PPI networks of organisms are shaped by evolution [202]. Indeed, almost all algorithmic solutions to the PPI network archaeology problem are based on explicit models of network growth (variations on the duplication–divergence principle), and take the form of parsimonious inference frameworks [187, 188, 202]; greedy local searches informed by models [83, 138, 139, 164]; or maximum likelihood inference of approximative [256], graphical [66], and Bayesian [111] models of the networks' evolution.

Less obvious is the fact that a second body of work, rooted in information theory and computer science, also makes the statement that growth processes can generate the history of real complex networks. This second strand of literature [32, 144, 146–148, 205, 219, 258] focuses on temporal reconstruction problems on tree-like networks generated by random attachment processes [16, 64]. It has led thus far to efficient root–finding algorithms with theoretical accuracy guarantees [32, 144, 205, 219], and to approximative reconstruction algorithms on trees [147, 148, 258]. Applying any of these algorithms to a real network amounts to assuming that growth processes—here random attachment models—are likely generative models.

The goal of the current paper is to directly investigate this notion of random *attachment* processes as generative models of the histories of networks, from the point of view of Bayesian statistics and hidden Markov processes [111]. Our contribution is threefold. One, we give a latent variable formulation of the network archaeology problem, in the context of a generalization of the classical preferential attachment model [5, 16, 128, 200]. We derive the full inference procedure for this model, including a sampling algorithm for its posterior distribution, as well as an efficient approximation thereof. Two, we establish the extent to which complete history reconstruction is possible, and, in doing so, identify a phase transition in the quality of the inferred histories (i.e., we find a phase where reconstruction is impossible, and a phase where it is achievable in large networks). Three, we demonstrate with numerical experiments that we can extract temporal information from real, static complex networks. We conclude by listing a number of important open problems.



(a) True arrival times (b) Posterior distribution

FIGURE 4.1 – Reconstructing the history of a growing network. (a) An artificial network generated by our generalization of the preferential attachment model (with parameters $\gamma = -1.1, b = 0.9, T = 50$, see main text). Since the network is artificial, its true history—i.e., the time of arrival of its edges in time—is known. The width and color of edges encode this history ; older edges are drawn with thick, dark strokes, whilst younger edges are drawn using thin, light strokes. The age of nodes is encoded in their radius. Our goal is to infer these times of arrival as precisely as possible, using the network structure as our only input. (b) Expected time of arrival, computed with 10^5 samples of the posterior distribution of the model. The correlation of the inferred and real history equals $\rho = 0.81$ (see Materials and Methods for details).

4.5 Results

4.5.1 Bayesian network archaeology

A network G generated by a growth process is, by construction, associated with a *history* X, i.e., a series of events that explains how G evolved from an initial state G_0. We consider the loosely defined goal of reconstructing X, using the structure of G as our only source of information (see Fig. 4.1). Formally, this is an estimation problem in which the history X is a latent variable, determined by the structure of the network. The relationship between the network and its history is expressed using Bayes' formula as

$$P(X|G,\theta) = \frac{P(G|X,\theta)P(X|\theta)}{P(G|\theta)} \,, \tag{4.1}$$

where it is assumed that the parameters θ of the growth process can be estimated reliably and separately.

To correctly define the probabilities appearing in (4.1), we first separate histories in two categories : Those that are *consistent* with G, and those that are not. We say that a history is consistent with a network if it has a non-zero probability, however small, of being the true history of the network. The likelihood $P(G|X,\theta)$ thus acts as a logical variable that enforces

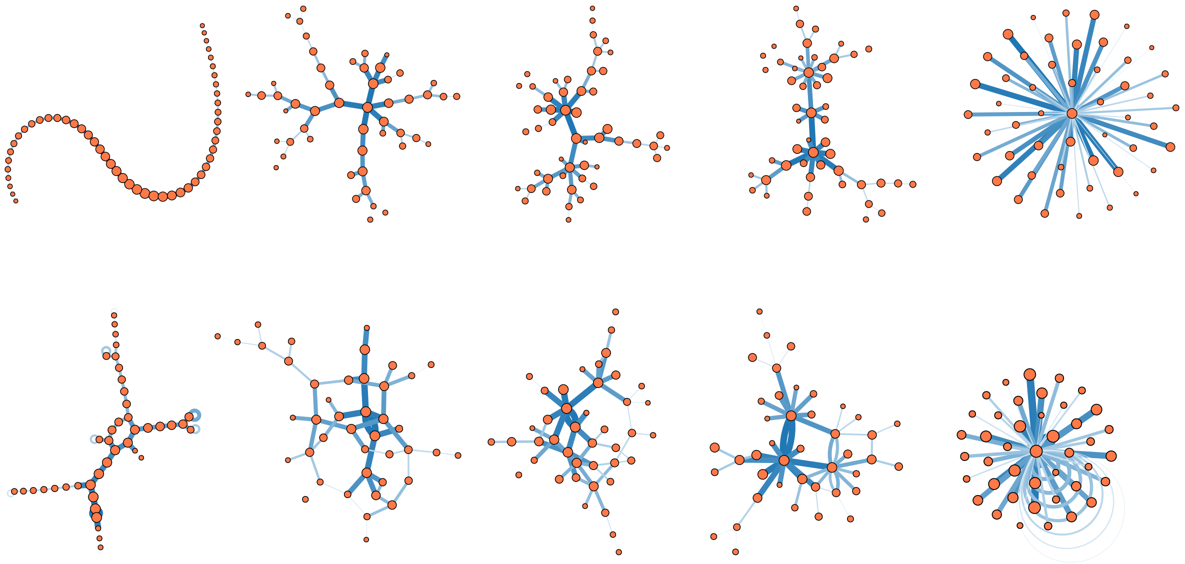


FIGURE 4.2 – Examples of networks generated by the process with $b = 1$ (top row) and $b = 0.75$ (bottom row), and $\gamma \in \{-10, -1, 0, 1, 10\}$ (from left to right). As in Fig. 4.1, the color and size of an edge indicates its arrival time, and likewise for the nodes.

this consistency : It is equal to one if and only if the history X is consistent with the observed network G, and it is equal to zero otherwise. A complete specification of the probabilities appearing in (4.1) is obtained upon choosing a growth process : This fixes the prior $P(X|\theta)$, as well as the evidence $P(G|\theta)$, because it is a sum of $P(X|\theta)$ over all histories consistent with G.

In this latent variable formulation of the network archaeology problem, reconstructing the past amounts to extracting information from G via the posterior distribution $P(X|G,\theta)$. Doing so is not as straightforward as it first appears. The posterior distribution may be heavily degenerated, or even uniform over the set of all histories consistent with G [146] (see Supplementary Information). Therefore, a useful and attainable goal cannot be to find the one true history $\tilde{X}(G)$ of G, because this history is often not identifiable. It turns out that another inference task, with a subtly different definition, is both challenging and achievable : That of finding a history as correlated with the ground-truth $\tilde{X}(G)$ as possible. We henceforth adopt this maximization as our inference goal.

4.5.2 Random attachment model

For the sake of concreteness, we will discuss of network archaeology in the context of a specific growth process. We use a variant of the classical PA model that incorporates both a non-linear attachment kernel [128] and densification events, i.e., attachment events between existing nodes [5, 129, 200].

In this model, a new undirected edge is added at each time step, starting from some initial network G_0, which we choose to be a single edge. With probability $1-b$ the edge connects two existing nodes, and it connects an existing node to a new node with complementary probability b. Whenever an existing node is involved at time $t+1$, it is chosen randomly with probability proportional to $k_i^\gamma(t)$, where $k_i(t)$ is the degree of node i at time t, and γ is the exponent of the attachment kernel.

The parameter $b \in [0,1]$ controls the density, and $\gamma \in \mathbb{R}$ controls the strength of the rich-get-richer effect (see Fig. 4.2). We refer to these parameters collectively with $\theta = (\gamma, b)$. We recover the classical PA model with $(\gamma = 1, b = 1)$, the random attachment model with $(\gamma = 0, b = 1)$ [64], and an undirected version of the Krapivsky-Redner-Leyvraz generalization [128] if γ is free to vary and $b = 1$. The model technically generates multigraphs for any b smaller than one, although the number of redundant edges and self-loops is vanishing for all b in the large network limit when $\gamma < 1$. It is thus a reasonable model of multigraphs, but also a good approximation of large, sparse networks, with few or no redundant edges and self-loops.

4.5.3 Inference algorithms

According to the model, every event marks the arrival of precisely one new edge. This allows us to represent histories compactly as an ordering of the edges of G in discrete time $t = 0, \ldots, T-1$, an arbitrary time-scale defined in terms of events [100]. Estimating the history then amounts to estimating the arrival times $\tau_{\tilde{X}}(e)$ of the edges $e \in E(G)$ in the ground-truth history $\tilde{X}$.

A good estimator $\hat{\tau}(e)$ of the arrival time of edge e is the posterior average :

$$\hat{\tau}(e) = \langle \tau_X(e) \rangle = \sum_{X \in \Psi(G)} \tau_X(e) P(X|G,\theta) \,, \tag{4.2}$$

where $\Psi(G)$ is the set of histories consistent with G. This estimator effectively combines histories, overcoming the degeneracy of the posterior distribution. It is straightforward to show that it minimizes the expected mean-squared error (MMSE) on $\tau_{\tilde{X}}(e)$, and we therefore refer to it as the MMSE estimator of the arrival time. Unfortunately, calculating the complete set of MMSE estimators $\{\langle \tau_X(e) \rangle\}$ is costly, because there are far too many histories consistent with networks of even moderate sizes. Hence, we resort to approximations.

We consider algorithms that fall in two broad categories : Sampling methods, whose goal is to evaluate (4.2) directly, but also topological methods that only rely on the structure of G to make predictions, forgoing explicit knowledge of the posterior distribution $P(X|G,\theta)$. Methods in the first category are the most accurate, but they are also more computationally demanding than, say, a simple sort of the nodes based on the degree.

With sampling methods, our goal is to a set of generate random histories $\{x_i\}_{i=1,\ldots,n}$ to approximate averages taken over $P(X|G,\theta)$. The most important property of this distribution is that its support $\Psi(G)$ only contains histories describing a growing network that is connected at all steps [the (γ,b) generalization of PA never generates disconnected components]. This constraint makes sequential importance sampling [111, 141, 241] an ideal choice, because it can explicitly enforce connectedness of the histories. Importance sampling relies on the transformation

$$\begin{aligned}\langle f(X)\rangle_P &:= \sum_{X\in\Psi(G)} P(X|G,\theta)f(X)\\ &= \sum_{X\in\Psi(G)} \frac{P(X|G,\theta)f(X)}{Q(X|G)}Q(X|G)\\ &\approx \frac{1}{n}\sum_{i=1}^{n}\frac{P(x_i|G,\theta)f(x_i)}{Q(x_i|G)},\end{aligned}\tag{4.3}$$

where the n samples $\{x_i\}_{i=1,\dots,n}$ are drawn i.i.d. from a proposal distribution $Q(X|G)$ different from $P(X|G,\theta)$. The idea is to use the proposal distribution to harness known properties of the target distribution—here the connectedness of growing networks. With efficiency in mind, we opt for a random recursive enumeration of the edges of G, associated with probabilities $Q(x_i|G)$ that can be easily computed on the fly, at no additional computational cost (see Materials and Methods). This proposal distribution—known as snowball sampling [135]—generates connected histories in linear time, i.e., in $O(|E| \times k_{\max})$ steps where $k_{\max}$ is the maximal degree.

With topological methods, our goal is to obtain a baseline performance, but also to investigate just how much information can be extracted without using sophisticated sampling methods. We consider two methods of linear complexity in the number of edges in the network. Both methods generate rankings of the nodes, and we use these rankings to induce a ranking on the edges, declaring ties when a pair of edges cannot be ordered. First, since age correlates with the degree in random attachment processes [4, 16], we consider a simple ordering of the edges by the degree of their nodes. Second, because edges closer to the center of a network are more likely to be old [33], we also use a method that harness the network's structure : the onion decomposition (OD) [101], a refinement of the k-core decomposition algorithm [18, 217] that splits the network into layers, from peripheral leaves to central nodes. A closely related "peeling" method has been recently introduced in Ref. [147] to tackle the archaeology problem in the case ($\gamma = 1$, $b = 1$).

4.5.4 Inference on artificial trees

We begin by testing the inference algorithms on trees drawn from the generative model itself (i.e., we set $b = 1$ and consider that γ is a free parameter). We compute the quality of a reconstruction using the Pearson product-moment correlation of the estimated arrival times and the ground-truth.

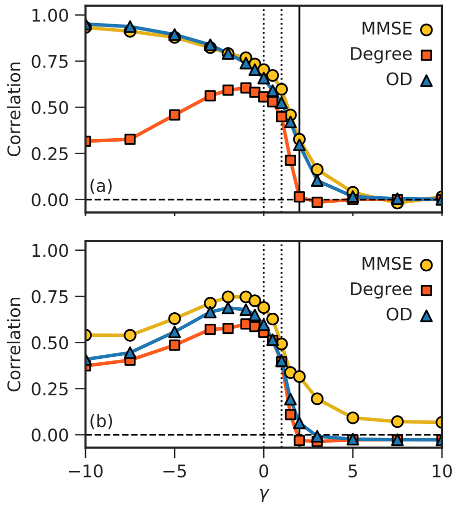


FIGURE 4.3 – Average correlation attained by the minimum mean-squared error (MMSE) estimators and two efficient methods based on network properties (a degree based method and the onion decomposition [101]), on artificial networks of $T = 50$ edges generated using our generalization of the preferential attachment model. (a) Phase transition for tree-like networks generated with $b = 1$. The special points corresponding to the uniform attachment model and the classical preferential attachment model are indicated with dotted vertical lines, at $\gamma = 0$ and $\gamma = 1$. The point where infinitely large networks fully condensate is shown with a solid vertical line at $\gamma = 2$ [128]. (b) Typical diagram for networks with cycles, here at $b = 0.75$. Each point is obtained by averaging the correlation obtained on 25 different network instances. We use $n = 5 \times 10^6$ samples to evaluate the MMSE estimators and a seed bias of $\alpha = 5$ (see Materials and Methods).

The average achieved correlation is shown as a function of the attachment kernel γ in Fig. 4.3 (a), on small networks ($T = 50$). We distinguish two regimes based on the performance of the degree estimators : The regime $\gamma > 0$, characterized by skewed distributions of degrees, and the homogeneous regime $\gamma < 0$. The three methods behave similarly in the former regime : They first yield a relatively large correlation at $\gamma = 0$, and their quality then quickly plummets with growing γ, ultimately converging to a null average correlation for sufficiently large values of γ. Throughout this transition, the MMSE estimators remain slightly superior to the OD estimators, and they both outperform the degree estimators by a significant margin. In contrast, the gap between methods is much larger in the homogeneous regime. While the quality of the OD and MMSE estimators increases with decreasing γ, the correlation achieved by the degree estimators goes in the opposite direction and shrinks with γ, eventually reaching 0 (not shown).

A better numerical portrait of the phase transition is shown in Fig. 4.4 (a), where we apply the efficient and nearly optimal OD method to increasingly larger networks. We find that for most values of $\gamma > 1$, the average correlation attained by the OD decreases as $T^{-\delta(\gamma)}$ with $\delta(\gamma) > 0$. If γ is close enough to 1, however, we do not observe any variations. This

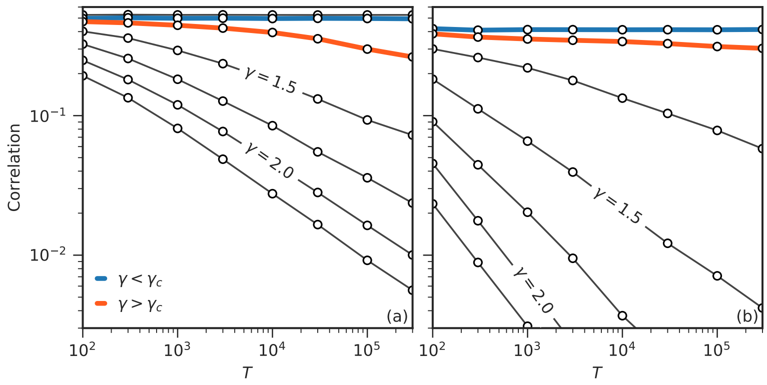


FIGURE 4.4 – Detail of the phase transition, through the lens of the OD. (a,b) Finite-size scaling analysis of the phase transition of Fig. 4.3. The average correlation attained by OD is shown against the size of networks generated with **(a)** $b = 1$ and **(b)** $b = 0.75$, for $\gamma \in \{1.00, 1.10, 1.25, 1.50, 1.75, 2.00, 2.25\}$ (from top to bottom). Curves that lie over (orange) and below (blue) the observed critical γ_c are highlighted.

suggests that non-trivial inference is possible in infinite networks, and signals a potential phase transition at some $\gamma_c > 1$.

4.5.5 Inference on artificial networks with cycles

It is clear that trees offer an easier challenge than general networks, because long-range loops (i) drastically increase the number of histories consistent with G, and (ii) introduce uncertainties in the ordering of large subsets of edges. To get a better understanding of the inference process, we therefore repeat the above numerical experiments on more general networks that include cycles, generated with $b < 1$. The outcome of our experiments are summarized in Fig. 4.3 (b) and Fig. 4.4 (b).

Allowing for cycles leads to three notable differences. First, we find that perfect reconstruction is no longer possible in the $\gamma < 0$ regime. Second, the separation between the MMSE estimators and the topological methods (OD, degree) becomes more pronounced for all γ. Third and finally, the transition becomes sharper and it occurs at a lower value of $\gamma_c(b)$; notice the much sharper decline in Fig. 4.4 (b).

4.5.6 A different task : root–finding

Inferring the complete history of a network is only one of many possible problems that fit within the Bayesian formulation of network archaeology. Any other temporal inference task may be attacked with the same set of tools. Let us therefore treat one such as the problem, as an example of the versatility of the framework : That of finding the root—the first edge—of G [219].

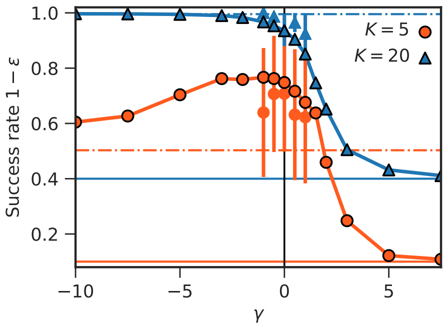


FIGURE 4.5 – Success rate of the root–finding algorithms, with sets $\mathcal{R}$ of sizes $K=5$ and $K=20$ on artificial networks of $T=50$ edges. The results of OD are shown with solid lines and closed symbols, while the sampling results are shown with open symbols and error bars of one standard deviation. The horizontal solid lines show the accuracy of randomly constructed sets $\mathcal{R}$ (no information retrieval), while the horizontal dotted lines shows the expected success rate in the limit $\gamma \to -\infty$. Sampling is not nearly as efficient as using OD for this problem, so we only compute them for $\gamma \in \{-1, -\frac{1}{2}, 0, \frac{1}{2}, 1\}$, with $n = 10^6$ samples. More samples are required to evaluate the set $\mathcal{R}$ correctly, which explains the large variance and poorer results of $K=5$.

In line with Refs. [32, 144, 205], we can give a solution to this problem in terms of sets : We define a procedure that returns a set $\mathcal{R}$ of $K(\varepsilon)$ edges, and guarantees that it will contain the first edge with probability $1-\varepsilon$. The size K depends on the acceptable error rate ε; larger sets cast a wider net, and are therefore more likely to contain the root.

To compute $\mathcal{R}$, we use a strategy based on the marginalization of the distribution $P(X|G,\theta)$. We first obtain the probability $P[\tau_X(e)=0]$ that an edge e is the first, for each $e \in E$, via

$$P[\tau_X(e)=0] = \sum_{X\in\Psi(G)} \mathbb{I}\left[\tau_X(e)=0\right] P(X|G,\theta)\,, \tag{4.4}$$

where $\mathbb{I}[S]$ is the indicator function, equal to 1 if the statement S is true, and to 0 otherwise. We then define $\mathcal{R}$ as the set formed by the K edges that have the largest posterior probability $P[\tau_X(e)=0]$. This probability is written in the form of (4.3) and it is therefore amenable to sampling. For comparison, we also infer the root with the much faster onion decomposition, by constructing $\mathcal{R}$ with the K most central edges (with ties broken at random).

The accuracy of the resulting algorithms is shown as a function of γ in Fig. 4.5. We distinguish, again, two main phases : Accurate reconstruction is possible in the strongly homogeneous regime $\gamma \ll 0$, but the success rate diminishes with growing γ, reaching a non-informative limit in the regime $\gamma \gg 0$. When $K \ll T$, the accuracy peaks at some value of $\gamma < 0$, before slowly decreasing to reach a smaller—but still informative—rate of success.

4.5.7 Application : The social network of a research institution

The end goal of network archaeology is to uncover temporal information from real temporal networks not *explicitly* generated by any growth process. We close this result section with an empirical demonstration that this goal is indeed achievable, here using the giant component of a growing social network, captured by email sent within a large European research institution [184] (see Materials and Methods).

The results of our analysis are summarized in Fig. 4.6, where we show the true history of the network, and compare the inferred histories with this ground-truth. We find estimated histories with positive correlations of $\rho_{\text{degree}} = 0.388$, $\rho_{\text{OD}} = 0.409$ and $\rho_{\text{MMSE}} = 0.615$. The network contains a dense connected core which is correctly placed first by all methods (the largest central nodes). More errors are made in ordering the periphery, because it contains fewer discriminating features (i.e., it is sparse in temporal information). Confirming that the (γ, b) generalization of PA is a good description of the observed network, we find that its degree distribution passes the standard test of goodness of fit of Ref. [42].

4.6 Discussion

There are two main themes to this paper : The accuracy of network archaeology on artificial networks—a mathematical problem—and network archaeology as a generative model for real systems—the related statistical tool. Let us now explore these topics in turn, beginning with the mathematical aspect of the problem.

4.6.1 Of information and phase transitions

The structure of a network generated by a growth model encodes its history. In the Results section, we have shown that this history can be recovered to varying degrees of accuracy, depending on the parameters of the generative model. These variations in accuracy, we now argue, are solely attributable to changes in the abundance of *equivalent edges*, i.e., edges that can never be ordered because they are topologically indistinguishable. Let us investigate how these classes of edges come about in our generalization of PA.

Model phenomenology

Let us first focus on the special case $b = 1$ and $\gamma \in \mathbb{R}$, thoroughly analyzed in Ref. [128]. This model has many known phases (see Fig. 4.2), characterized by different degree distributions. In the limit $\gamma \to -\infty$, the model generates long paths, where every node has degree 2 except for the two end-nodes, of degree 1. For all negative values of γ, the precise form of the degree distribution is not known, but it is clear that the model favors degree homogeneity. When $\gamma = 0$, the degree distribution is geometric, of mean 2 (since we recover the uniform attachment model [64]). In the interval $0 < \gamma < 1$, the degree distribution takes the

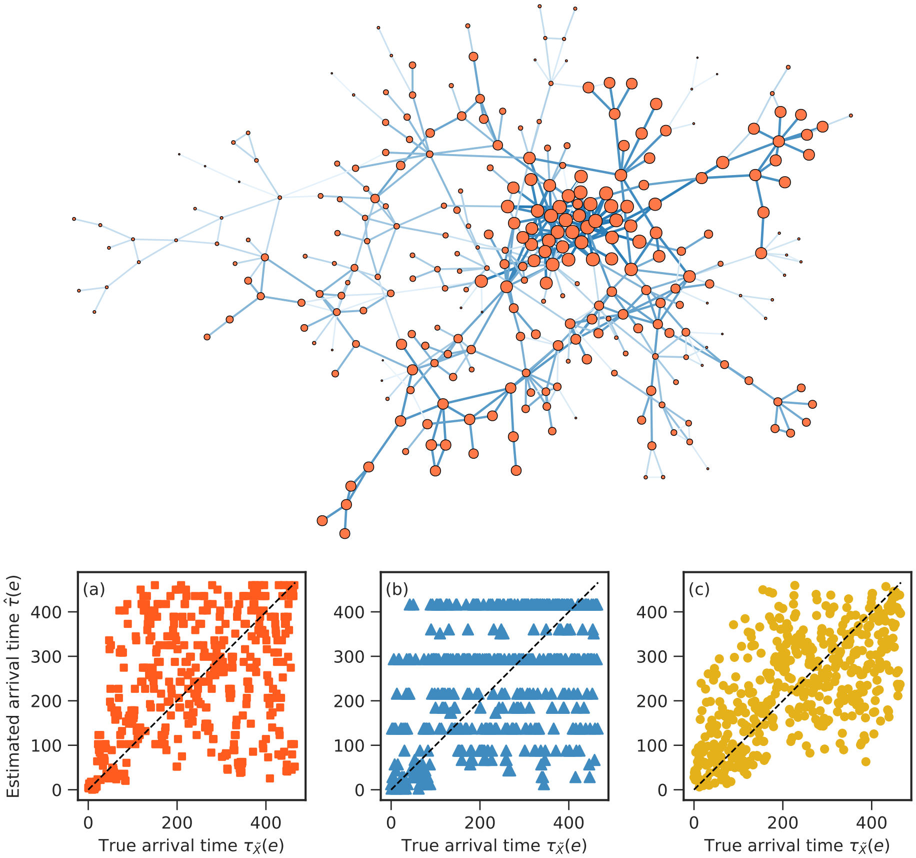


FIGURE 4.6 – Application of the inference framework to the giant component of the social network of a large European research institution. (top) Nodes ($n = 294$) represent the members of the institution [184], and edges ($T = 406$) represent strong social ties, as captured by numerous email exchanges. Older edges are drawn with thick, dark strokes, whilst younger edges are drawn using thin, light strokes. The radius of a node is proportional to its age. We find that this network is best modeled by $\hat{b} = 0.63$ and $\hat{\gamma} = 0.07$. This pair of parameters leads to a KS–statistic of $D^* = 0.040$ and a p-value $P[D > D^*] = 0.24$ under the random model, signaling a good fit (see Materials and Methods). (bottom) Inferred arrival time versus true arrival time. A perfect result is shown for reference (diagonal dashed line). The attained correlations are : **(a)** $\rho_{\text{Degree}} = 0.388$, **(b)** $\rho_{\text{OD}} = 0.409$, and **(c)** $\rho_{\text{MMSE}} = 0.615$. The latter is evaluated with $n = 10^4$ samples.

form a stretched exponential, with an asymptotic behavior fixed by γ. At precisely $\gamma = 1$, the attachment kernel becomes linear and the networks scale-free : The degree distribution follows a power-law of exponent -3. In the interval $1 < \gamma < 2$, the networks *condensate* in a rapid succession of phase transitions at $\gamma_m = (m+1)/m$ for $m \in \mathbb{N}^*$. When $\gamma > \gamma_m$, the number of nodes of degree greater than m becomes finite. As a result, an extensive fraction of the edges aggregates around a single node—the condensate—and this fraction grows with increasing γ [181]. The condensation is complete at $\gamma = 2$, where the model enters a winner-takes-all scenario characterized by a central node that monopolizes nearly all the edges, with high probability.

Scalable inference and the no-recovery phase

The appearance of a condensate from $\gamma > 1$ onwards gives a nice qualitative explanation of the phase transition observed in Figs. 4.3–4.5. The important insight is that the condensate significantly limit our ability to infer the history of a network, due to the fact that it scales with the system size. When an edge attaches to the condensate, the temporal information it carries is destroyed, because all the edges of a star-graph are structurally equivalent, and therefore not orderable. Hence, the diminishing correlation of the estimators in the regime $\gamma > 1$ can be seen as arising from the creation of a larger and larger set of equivalent edges. When the condensation phenomenon is strong enough at some critical γ_c, the models enter a no-recovery phase where all temporal information is lost.

The position of γ_c of the threshold is of theoretical interest. It is certainly no greater than 2, because the star-like graphs of this regime—i.e., full condensates—are completely unorderable. If we are able to find positively correlated histories beyond $\gamma = 2$ in Fig. 4.3, it is only because the system is small. Scale the network up and our predictive power vanishes, see Fig. 4.4. But while it is clear from the phenomenology of the generative model that the transition lies at some $\gamma_c \leq 2$, its exact location is harder to pinpoint.

The difficulties stem from the fact that we can only inspect the transition through the lens of a good, but suboptimal algorithm. Even if the finite-size scaling analysis of Fig. 4.4 tells us that the OD fails at some γ_c^{OD}, it is possible that a better method will work until a true no–recovery threshold $\gamma_c \geq \gamma_c^{\mathrm{OD}}$ is reached. That being said, we can reach a number of conclusion from the imperfect analysis of Fig. 4.4. First, the figure tells us that $\gamma = 1$ is only special in that it marks the end of the regime where there are no obvious information sinks. Scalable inference appears possible well above the appearance of a condensate. Second, the figure suggests that the OD fails at $11/10 < \gamma_c^{\mathrm{OD}} < 5/4$ in the infinite network limit, because it achieves an average correlation that goes to zero with growing T above γ_c^{OD}, while it converges to an average correlation greater than 0 at γ_c^{OD}.

Whether an optimal algorithm will behave similarly remains an open question. Based on the effect of the condensate on the remainder of the network, we conjecture that γ_c is aligned with one of the transitions $\gamma_m = (m+1)/m$ of the generative model, for $m \leq 9$ (from Fig. 4.4). The most likely value is $\gamma_c = 3/2$, because the networks generated by the model change *drastically* at this point. Above $\gamma = 3/2$, they comprise of nodes of degree 2 or less that can—and do—cluster around a highly attractive central node (encoding next to no temporal information). Below $\gamma = 3/2$, the networks contains infinitely many nodes of degree 3 that must necessarily organize in large scale structures (that encode temporal information).

Nearly perfect inference

While the overabundance of nodes of low degree leads to the onset of a no–recovery phase in the regime $\gamma \gg 0$, these nodes have the opposite effect at the other extreme $\gamma \ll 0$. In the presence of extreme homogeneity, the networks are effectively grown as a random path, where all nodes are of degree two except the two end-nodes, of degree one. All edges are then clearly ordered up to a symmetric flip around the center of the path. Standard concentration inequalities tell us that the time of arrival of any edge can be identified up to an uncertainty of vanishing size in the large T limit. Near perfect inference is therefore trivial : Peeling the path symmetrically from both sides yields a close approximation of the arrival time of every edge.

Effect of cycles

The above conclusions explicitly rely on the fact that the networks are trees, i.e., that $b = 1$. It turns out that many of these conclusions carry over to the case $b < 1$, as highlighted by the similarity of the numerical results in Figs. 4.3–4.4. There are essentially two areas where allowing for cycles brings notable differences : Near-perfect inference becomes impossible when $\gamma \ll 0$, and the correlation decreases much faster with growing γ.

The disappearance of the near perfect inference phase in the regime $\gamma \ll 0$ is imputable to the appearance of long-range connections. These connections close dangling paths and erase all temporal information along them. The net result is that the limit $\gamma \to -\infty$ actually poses a hard challenge for any $b < 1$, mirroring the limit $\gamma \to \infty$. The near perfect inference phase is thus highly uncharacteristic of the general model.

If the correlation diminishes more abruptly when $b < 1$, it is because cycles, self-loops and parallel edges—all allowed motifs in the regime $b < 1$—accentuate the condensation phenomenon. A typical network realization in the super-linear regime $\gamma > 1$ with $b < 1$ comprises of : Many unorderable self-loop centered on the condensante ; a number of parallel edges connecting high degree nodes ; and star-like node arrangements around high-degree nodes. The information sinks of the case $b = 1$ are thus both larger and better interconnected.

4.6.2 On the quality of the inference methods

Our numerical results suggest that even though the naive degree-based approach works relatively well in the regime $\gamma > 0$, the MMSE estimators and the onion decomposition actually perform much better across the board. Of these two methods, the correlation attained by the MMSE estimators is perhaps the least surprising; they after all build on an explicit knowledge of the growth model. In fact, we can write the average correlation of an inferred history Y with the ground-truth as $\langle\rho(X,Y)\rangle \propto \rho(Y,Z)$ where the average on the left-hand side is taken over $P(X|G,\theta)$ and where Z is the MMSE history (see Supplementary Information). Hence, anything short of using the MMSE estimators yields worst results on average.

The result that needs explaining, then, is the excellent average correlation achieved by the OD. A simple combinatorial argument can explain this performance. Notice that for a network G, the vast majority of histories $X \in \Psi(G)$ place central edges at the beginning, and peripheral edges at the end [33, 219] (both in trees and general networks). This is a consequence of the fact that there are many more consistent ways of enumerating the graph G starting from its center than from its periphery. The net result is that the MMSE estimators of the central edges are heavily skewed towards early arrival time, while that of the edges in the periphery are skewed towards later times (see (4.2)). In other words, the separation in layers uncovered by the OD is contained directly in the posterior distribution—the onion decomposition is good *because* it approximates the optimal MMSE estimators. This is a useful connection, because the OD is far more efficient than sampling : The OD returns its *final* estimators in $O(|E| \times \log|V|)$ steps [101], whereas a single *sample* is generated in roughly as many steps by the importance sampling algorithm.

4.6.3 New insights on related analyses

At this point, we ought to discuss important connections with related work on root–finding [32, 144, 205, 219] and complete history inference [146, 147], on random and preferential trees [i.e, for the models $(\gamma, b) = (0,1)$ and $(1,1)$]. These previous analyses establish optimal algorithms to find *node* orderings; so let us define the operator $\tilde{\tau}_X(v)$, identical to $\tau_X(e)$ on vertices.

The most comprehensive root–finding method is put forward in Ref. [32]. Their strategy is to compute the number $\varphi(v) = |\{X|\tilde{\tau}_X(v) = 0\}|$ of histories rooted on v, and to return the K nodes with the largest $\varphi(v)$. They show that this algorithm can be employed to construct sets of constant size that contains the root with a fixed error rate $\varepsilon < 1$ as T goes to infinity, and that the case $\gamma = 0$ is easier than the case $\gamma = 1$ (smaller sets are needed to attain the same error rate ε). Our results (Fig. 4.5) corroborates these observations and put them in the broader context of Bayesian inference with general γ (a generalization suggested in Ref. [32]).

For example, notice that $\varphi(v)$ is in fact proportional to the posterior probability :

$$\varphi(v) \propto \sum_{X \in \Psi(X)} \mathbb{I}[\tilde{\tau}_X(v) = 0] P(X|G,\gamma) \equiv P[v \text{ is first}|G,\theta], \tag{4.5}$$

when the distribution $P(X|G,\gamma)$ is uniform. Thus the estimators of Refs. [32, 205, 219] can be in fact seen as outputting the K first maxima of the marginal distribution for the first node of G, assuming a uniform posterior distribution. The general estimator appearing in (4.4) accounts for general parametrization, and point the way to obvious extensions to more complicated root graphs [144].

Turning to related work on complete history recovery, we note that the recent analysis of Ref. [147] also relies on a peeling algorithm, almost isomorphic in its action to the onion decomposition [101]; it is shown in this reference that the algorithm performs extremely well on scale-free trees ($\gamma = 1$ and $b = 1$), in line with our analysis. Importantly, the notion of inference quality used in Ref. [147] is different from ours, because the authors point out the fact that there is a trade-off between precision and density (number of ordered pairs of nodes). In particular, when the peeling algorithm is allowed to withhold judgment on contentious pairs of nodes, one obtains nearly perfectly accurate estimates, at the cost of a small estimate density. Our analysis showcases another aspect of the power of peeling–type algorithms : They are also nearly optimal—i.e., as effective at extracting information as the best estimators—when all pairs must be ordered.

4.6.4 Generative modeling of real networks

Our generalization of PA makes sense as a model for the email network investigated in the result section (see Fig. 4.6). New individuals joining the network must necessarily attach *somewhere*—a phenomenon that we model through the initial edge of every node. Furthermore, edges do appear after this initial event—which we model with $b < 1$.

The estimated parameters are consistent with the fact that the email network is overlaid on a social system. First, we find that attachment is far from preferential (with an estimated kernel exponent of $\hat{\gamma} \approx 0.07$). This is in line with the observation that social capacities are limited, as famously argued in the anthropological theory known as Dunbar's number [65]. The degree distribution cannot be too skewed, because there are constraints on the maximal number of connections that can be entertained simultaneously by any given individual. Second, the estimated ratio $\hat{b} \approx 0.62$ signals that existing nodes share many edges—the network, while sparse, is definitely not a tree. This can be seen as a consequence of the fact that social networks are dense webs of connections, most of which are not formed at attachment (they are instead formed later through, e.g., triadic closure [88]).

A reliable indicator that the model is suitable for the investigated system is the ordering of the inference results of Fig. 4.6. Paralleling the results on artificial networks in the regime

$\gamma = 0.07$ with cycles, we find that the MMSE estimators are the best, but they are closely followed by OD, with the degrees behind. This tells us that temporal information is encoded in the network in a way that is consistent with the generative model, and shows that (some) temporal information can indeed be extracted from real network structure.

4.6.5 Looking ahead : challenges and generalizations

The opportunities brought about by network archaeology are tantalizing ; in bioinformatics alone—the only field where it has found widespread adoption thus far—, network archaeology with the divergence-duplication models has already yielded insights into the past states of real PPI networks [164, 202] and, e.g., improved on network alignment [66, 75]. Generalization to models that are relevant to social and technological networks will allow us to answer new questions about the past of static systems, and to improve on network analysis techniques [91].

The present paper provides a simple framework to carry this program forward. With a model specified as a Markov chain, importance sampling provides weighted histories that can be aggregated as MMSE estimators to yield optimal estimators of the true history of a network. Different models will require different proposal distributions, but the prescription is almost automatic. Drawing from a background network will always work, as long as histories leave tangible traces (i.e., there are no deletion events [188]).

That being said, our analysis is of course not complete, and leaves a number of important theoretical and computational problems open. First, while we have provided compelling evidence for the existence of a scalable inference and no-recovery phases (in the general model $b \leq 1$), we have not pinpointed the location $\gamma_c(b)$ of the transition that separates them ; our analysis suggests that it lies at some rational value $\gamma_c = (m+1)/m$ for $m \leq 9$ (probably at $\gamma_c = 3/2$ when $b = 1$.) Second, we have shown that the MMSE estimators are optimal in the sense that they maximize the correlation averaged over the posterior distribution. This invites the question : Do they also satisfy other notions of optimality ? Finally, we have argued for using the OD [101] in large networks because importance sampling is not scalable, at least not for the purpose of network archaeology. This substitution will most likely not work with different models, because it is based on a close correspondence between the posterior averages of arrival times, and the order of peeling of a network. As a result, the next step will be to derive efficient approximation methods that work with general models, to allow for flexible network archaeology. These methods will have to handle models specified as chains $P(X|\theta)$ with some arbitrary notion of consistency $P(G|X,\theta)$. The relaxation technique of Ref. [140] for permutation inference comes to mind ; but also dynamical variants of message-passing [153], perhaps in the spirit of Ref. [143].

4.7 Materials and Methods

4.7.1 Generative model

We consider a generative model that generalizes the classical preferential attachment model (PA) of Barabási-Albert [16]. The salient features of the generalization are a non-linear attachment kernel k^γ [128] and the possibility for new links to connect pairs of existing nodes.

This model generates sequences of undirected graphs $G_0, ..., G_{T-1}$, where G_{t-1} has one fewer edge than G_t. We summarize a particular sequence of graphs using the *history* of the generative process, i.e., a tuple $X = (e_0, ..., e_{T-1})$, where e_t is the edge added to the sequence at time step t.

Random growth events are resolved as follows. At time step t, we first draw a random node from V_t, the node set of G_t prior to any modifications of the graph's structure. Node i is selected with probability

$$w_i(\gamma, t) = \frac{k_i^\gamma(t)}{\sum_{j\in V_t} k_j^\gamma(t)}\ , \tag{4.6}$$

where $\gamma \in \mathbb{R}$ is the attachment kernel, and where $k_i(t)$ is the degree of node i at time t, *before* the new edge is added. We then complete the edge with a *new* node with probability b, and with an existing node with complementary probability $1-b$ (a *densification* event). When densification events occur, the second is selected randomly with weights given by (4.6). This model corresponds to an out of equilibrium Markov process, such that the prior distribution $P(X|\gamma, b)$ is merely a product of transition probabilities.

The posterior probability $P(X|G,\gamma,b)$ is obtained by conditioning on a labeled graph G. To do so, we first define the *consistency* of a history X with a graph G, denoted $X \Rightarrow G$. A history X is said to be consistent with G if the labeled graph G' with edge set $\{e|e \in X\}$ is isomorphic to G, and $P(X|\gamma,b) > 0$. It is easy to see that there are many different histories consistent with G, but that there is only one labeled graph G' associated with each X (there is a surjection of history onto graphs). Therefore, the probability $P(G|X,\gamma,b) = 1$ if and only if X is consistent with G, and it is normalized. Bayes formula allows us to write

$$P(X|G,\gamma,b) = \frac{P(X|\gamma,b)}{P(G|\gamma,b)}\mathbb{I}[X \Rightarrow G]\ , \tag{4.7}$$

where $\mathbb{I}[S]$ is the indicator function, equal to 1 if the statement S is true, and to 0 otherwise. Although not directly useful for estimation purposes, the probability $P(G|\gamma,b)$ is in principle defined. It is obtained by summation :

$$P(G|\gamma,b) = \sum_{X\in\Phi_T} P(X|\gamma,b)P(G|X,\gamma,b) = \sum_{X\in\Psi(G)} P(X|\gamma,b)\ , \tag{4.8}$$

where Φ_T is the set of all histories of length T, and where $\Psi(G) \subseteq \Phi_T$ is the set of histories consistent with G.

4.7.2 Inference task

Let $\tau_X(e) \in \{0,...,T-1\}$ denote the position of edge e in history X (also called its arrival time). Our goal is to give the best possible estimate of $\{\tau_{\tilde{X}}(e)\}_{e\in G}$, using only the structure of the labeled graph G, where $\tilde{X}$ is the real history of G (the ground-truth). By convention, we express both the estimators and the true history in the time scale $t = 0,\dots,T-1$ where $T = |E|$. Note, however, that the estimator $\hat{\tau}(e)$ of $\tau_{\tilde{X}}(e)$ need not be an integer, or distinct from other estimators [i.e., we allow $\hat{\tau}(e) = \hat{\tau}(e')$].

We quantify the quality of the estimators $\{\hat{\tau}(e)\}$ of the true arrival times using the Pearson product-moment correlation coefficient

$$\rho = \frac{\sum\limits_{e\in E(G)} \big(\tau_{\tilde{X}}(e) - \langle\tau\rangle\big)\big(\hat{\tau}(e) - \langle\tau\rangle\big)}{\sqrt{\sum\limits_{e\in E(G)} \big(\tau_{\tilde{X}}(e) - \langle\tau\rangle\big)^2}\sqrt{\sum\limits_{e\in E(G)} \big(\hat{\tau}(e) - \langle\tau\rangle\big)^2}}\,, \tag{4.9}$$

where $\langle\tau\rangle = (T-1)/2$, and where we have dropped the subscript for the sake of conciseness. The correlation takes values in $[-1,1]$, where $|\rho| = 1$ indicates a perfect recovery up to a time reversal, and where $|\rho| = 0$ indicates that no information is extracted from the graph at all In taking the sum, it is assumed that the edges are distinguishable (this matters for multigraphs with self-loops). The Pearson product-moment correlation has two useful properties adapted to the network archaeology problem. One, it is not affected by an arbitrary linear transformation of the timescales; this captures the fact that the timescales of the compared histories are in fact arbitrary. Two, it penalizes spurious ordering of events. If a graph contains no information on the ordering of a subset $S \subseteq E$ of the edges, creating a random ordering of the edges in S yields a worst outcome on average than attributing the average time of arrival λ of all edges in S, to all edges in S (see Supplementary Information).

4.7.3 Topological estimation algorithms

We regroup in this category the methods that rely only on the structure of G to make predictions, forgoing explicit knowledge of the posterior distribution $P(X|G,\theta)$. These methods all follow the same formula : We first rank the edges, based on some network property (degree, centrality, etc.), and then output these ranks as $\{\hat{\tau}(e)\}$, the estimated arrival times. Whenever the edges of a subset $S \subseteq E$ are indistinguishable, we give them the same rank $\lambda(S)$, reflecting our uncertainty of their true ordering. We chose a $\lambda(S)$ that preserves the overall average time of arrival $\langle\tau\rangle$; this yields $\lambda = t + (m+1)/2$, where $m = |S|$ and $t+1,...,t+m$ are the ranks that would have been assigned to the edges of S, had they been ordered. This choice is optimal in the sense that assigning any other rank to the edges of S would not reliably increase the overall correlation (see Supplementary Information). Furthermore, it has the added benefit that the Pearson correlation can then be computed directly as Spearman's rank correlation, if one considers the arrival time as the rank.

Degree-based estimation

Nodes that arrive earlier in the process have, on average, a larger degree [4, 16]. We use the degree of nodes to induce a ranking of edges as follows. Let $(k_e^{\text{low}}, k_e^{\text{high}})$ denote the degree of the nodes connected by edge e, with $k_e^{\text{low}} \leq k_e^{\text{high}}$. We rank edges in descending order with k_e^{high}, and break ties with k_e^{low}, when possible. The equalities that remain are declared as such. We then use these ranks as the estimated time of arrival.

Layer-based estimation

The onion decomposition (OD) generalizes the k-core decomposition. The method is based on the classical $O(|E| \times \log|V|)$ algorithm for the k-core decomposition [18, 101] : it peels networks by iteratively removing all nodes, starting from the lowest degree nodes. Nodes are assigned a coreness number equal to their degree when they are removed. Different from the classical algorithm, the OD treats each batch of removal as a separate layer ; this assigns both a coreness and a layer number to each node. We assume that nodes with the lowest coreness numbers appeared last and that, within a coreness class, the first removed nodes are the youngest. A simple modification allows the algorithm to order the *edges* : An edge is assigned to a pass as soon as one of its nodes is peeled away. All edges removed in the same pass are declared as tied. Because we effectively discard the coreness number to order edges and nodes, the OD is almost equivalent to the peeling algorithm of Ref. [147], that proceeds by iteratively removing the lowest degree nodes. The only difference is that nodes of low but different degrees are sometimes removed simultaneously in the OD [101].

4.7.4 Sampling algorithms

Estimators

A principled estimation of $\tau(e)$ must rely on the posterior distribution $P(X|G,\gamma,b)$. We use the minimum mean square error (MMSE) estimator of $\tau(e)$, given by

$$\hat{\tau}(e) = \langle \tau_X(e) \rangle_{P(X|G,\gamma,b)} = \sum_{X\in\Psi(G)} \tau_X(e) P(X|G,\gamma,b) \,, \tag{4.10}$$

since it maximizes expectation of the correlation a posteriori (see Supplementary Information). It is more robust than the other natural alternative, namely the mean maximum overlap (MMO) estimator, obtained by finding the time t that maximizes the marginal posterior probability $p_t(e) = \sum \mathbb{I}[\tau_X(e) = t] P(X|G,\gamma,b)$ over all t. Indeed, the maximum of $p_t(e)$ with respect to t is frequently not unique, because many pairs of edges are truly unorderable [147]. In the presence of the slightest perturbation, these equivalence classes collapse (or must be diagnosed using a costly post-processing step), leading to poor performances of the MMO estimator. In contrast, the MMSE estimators hardly change when the empirical estimation of $p_t(e)$ is inexact, leading to robust predictions.

Importance sampling

The support of the posterior distribution is far too large to handle analytically. This forces us to resort to sampling methods, and approximate $\hat{\tau}(e)$ as

$$\hat{\tau}(e) \approx \frac{1}{n}\sum_{i=1}^{n} \tau_{x_i}(e) \, , \tag{4.11}$$

where x_i is a random history of length T drawn from the posterior distribution.

In practice it is hard to sample from $P(X|G,\gamma,b)$, so we will prefer a transformation of (4.10). Given an easy-to-sample stochastic process that enumerates the edges of G with probability $Q(X|G)$, we express the MMSE estimator of $\tau(e)$ as

$$\begin{aligned}\hat{\tau}(e) &= \sum_{X\in\Psi(G)} \frac{\tau_X(e)P(X|\gamma,b)}{P(G|\gamma,b)}\frac{Q(X|G)}{Q(X|G)} \\ &= \frac{\langle \tau_X(e)\omega(X|G,\gamma,b)\rangle_{Q(X|G)}}{P(G|\gamma,b)} \, ,\end{aligned} \tag{4.12}$$

where $\omega(X|G,\gamma,b) = P(X|\gamma,b)/Q(X|G)$ is the sample weight. Equation (4.12) is then well approximated by

$$\hat{\tau}(e) \approx \frac{1}{nP(G|\gamma,b)}\sum_{i=1}^{n} \tau_{x_i}(e)\omega(x_i|G,\gamma,b), \quad \text{where } \{x_i\} \overset{i.i.d.}{\sim} Q \, ,$$

in the large n limit. But contrary to (4.11), this second formulation has a control parameter—the *proposal distribution* $Q(X|G)$—that can be used to simplify the sampling process and/or enhance convergence. This re-weighting scheme is known as the importance sampling method [11].

Snowball sampling

We use the proposal distribution $Q(X|G)$ to simplify the sampling procedure. A snowball sample is a random recursive enumeration of a graph, rooted at a randomly selected seed [70, 135]. More explicitly, given a random initial edge e_0 (the seed), define the *boundary* as the set of all non-enumerated edges that share at least one node with e_0. Then select an edge e_1 from the boundary uniformly at random, and update the boundary by adding all new edges reached with e_1, repeating the process for $e_2, e_3, \ldots$, until the graph is exhausted. Because this algorithm draws the edge from the observed graph G (the "background graph" [33]), it generates a history X that is necessarily consistent with G. Thus, the algorithm is efficient in the sense that it (a) always generates a plausible history without rejection, and (b) has a polynomial worst-time complexity of $O(|E| \times k_{\max})$, where $k_{\max}$ is the maximal degree.

The probability of generating a history X with the snowball sampling algorithm is given by

$$Q_{\mathrm{sb}}(X|G) = P(e_0|G) \times \prod_{t=1}^{T-1} \left[|\Omega_X(t)|\right]^{-1} \, , \tag{4.13}$$

where $\Omega_X(t)$ denotes the boundary at step t of history X, and where $P(e_0|G)$ is the distribution used to select the seed e_0. To bias the proposal distribution towards histories that begin with edges attached to high-degree nodes, we select the initial edge with probability proportional to $(k_1 \times k_2)^\alpha$, where α is a parameter, and where (k_1,k_2) are the degrees of the nodes connected by an edge. This better mimics the ground truth and leads to a faster convergence of the estimators. Because the histories X are constructed iteratively, the resulting sampling method is called a *sequential importance sampling algorithm* [141].

Normalization

In modifying the expression for the estimator, we have introduced an intractable normalization $P(G|\gamma,b)$. This issue can be side-stepped by noticing that the estimators $\langle\tau(e)\rangle$ must satisfy the sum

$$\begin{aligned}\sum_{e\in E(G)}\langle\tau(e)\rangle &= \sum_{e\in E(G)}\sum_{X\in\Psi(G)}\tau_X(e)P(X|G,\gamma,b)\\ &= \sum_{X\in\Psi(G)}P(X|G,\gamma,b)\sum_{e\in E(G)}\tau_X(e) = \binom{T}{2}\,, \end{aligned} \tag{4.14}$$

where the last equality follows from the normalization of $P(X|G,\gamma,b)$. Thus we compute $\hat{\tau}(e)$ up to a multiplicative constant and use (4.14) to set the scale of the estimators.

4.7.5 Parameter estimation

In writing the sampling methods, we have conditioned each estimator $\hat{\tau}(e)$ on (γ,b), because as we now show, the parameters can be estimated from a single, static, observed network G.

The ratio of the number of nodes to the number of edges

$$\hat{b}(G) = \frac{|V(G)|-2}{|E(G)|-1} \tag{4.15}$$

is an obvious estimator for b. This is due to the fact that the observed graph can be seen as a signature of $|E|-1$ i.i.d. Bernoulli trials of parameter b. Each edge beyond the first embodies a test of whether a new node should be added, and each node beyond the two initial nodes signals a success of the trial.

The exponent γ poses a harder challenge. Direct estimation involves complex a posteriori estimates of $P(\gamma|G,b) \propto P(\gamma|b)\sum_X P(X|\gamma,b)$, where $P(\gamma|b)$ is a Bayesian prior on γ. Neither maximizing this posterior distribution nor evaluating its average is easy, since both approaches require the calculation of an intractable sum (where γ appears as a continuous exponent). In particular, this prevents the use of known estimation methods based on time-resolved graphs [87, 113, 201], because too many estimates would have to be combined.

We instead opt for a simpler solution based on the degree distribution, namely the now standard Kolmogorov-Smirnov (KS) minimization approach of Ref. [42]. The KS–statistic of a pair of noisy distributions (P, Q) is given by the supremum of the difference of their cumulative distribution function (CDF), i.e.,

$$D(P, Q) = \sup_k |f_P(k) - f_Q(k)| \, , \tag{4.16}$$

where $f_P(k)$ is the CDF of P at point k. It is equal to zero when the compared distributions are identical. Given an empirical degree distribution $P(G)$ derived from an observed network G, we learn γ by minimizing the KS–statistic averaged over a set of n random degree distributions $\{Q^{(i)}(\gamma)\}_{i=1,\ldots,n}$ generated by the model of parameters $(\gamma, \hat{b})$. The minimum $D^*(G)$ can be found efficiently using Brent's method [204], since the average KS–statistic is convex. We use $n \gg 1$ random instances to compute D at each evaluation, to ensure that the resulting function is smooth enough for the optimizer to converge to its true minimum. When T is large, generating $n \gg 1$ networks can be costly. Therefore, in practice, we instead first compute the expected degree distribution of the model using mean-field equations, and then draw n finite samples from the resulting distribution. This approach is equivalent to—but much faster than—direct simulations.

We note in closing that the above framework provides a natural notion of goodness of fit for γ [42]. Indeed, the goodness of fit can be assessed by first generating few random degree distributions with the estimated parameters $(\hat{\gamma}, \hat{b})$, and then applying the complete testing procedure anew, using the random degree distributions as the reference empirical distributions. This provides a null distribution for D, which tells us whether $D^*(G)$ is an extreme value of the average KS–statistic or not. Following the standard, we assume that if $P[D > D^*(G)] > 0.1$, then the fit is good [42].

4.7.6 Experimental pipeline

In all numerical experiments, we randomize both the node labels and the order in which the edge list is passed to the algorithms, as to hide any implicit temporal information contained in the tags of the nodes and ordering of the edges. All posterior averages are evaluated with the true parametrization of the model, when the networks are artificial (Fig. 4.3 and 4.5). The parameters are estimated in experiments involving real networks.

4.7.7 email-Eu-core dataset

The datasets analyzed in the Result and Discussion sections consists of a series of anonymized emails, sent within a large European research institution [184]. It comprises 332 334 emails sent between 986 individuals, over a span of 803 days. To construct a single undirected growing network from a series of emails (directed edges), we consider that two individuals become connected as soon as they have both sent $n_c = 40$ emails to the other party. This filtration removes noise (e.g. institutional spam) and spurious interactions (e.g. transient contact) from the dataset. Our choice of threshold is motivated by the fact that the resulting network is not too small, yet sparse enough to be well-modeled by the (γ, b) generalization of PA. The results are robust to small changes in n_c, see Supplementary Information. We focus on the final giant component of the resulting network (294 nodes and 406 edges). To construct its history, we start from the end of the growth process, and out the largest component. We then run its history—the sequence of new edges—backward. When it absorbs smaller components (i.e., when more than one edge is removed from the component simultaneously in its reversed history), we consider that their edges appeared sequentially, as if they had been created at the time of the absorption.

4.7.8 Software

Implementations of the generative models and inference methods are available online at `github.com/jg-you/network_archaeology`.

4.8 Supplementary Material

4.8.1 Characterization of the generative model

The generative model of the main text is a generalization of the classical preferential attachment model [16]. The salient features of the generalization are a non-linear attachment kernel k^γ [128] and the possibility for new links to connect pairs of existing nodes [5, 62, 129, 200]. See the "Material and methods" section of the main text for a precise definition.

Degree distribution

Denote by $N_k(t)$ the number of nodes of degree k at time t and define the normalization $Z_\gamma(t) = \sum_k N_k(t)k^\gamma$. The evolution of $\{N_k(t)\}$ is approximately governed by the set of differential equations

$$\frac{dN_1(t)}{dt} = b - \frac{N_1(t)}{Z_\gamma(t)}[1 + (1-b)] \, , \tag{4.17a}$$

$$\frac{dN_k(t)}{dt} = [1 + (1-b)]\frac{\left[N_{k-1}(t)(k-1)^\gamma - N_k(t)k^\gamma\right]}{Z_\gamma(t)} \, , \tag{4.17b}$$

$$\frac{dZ_\gamma(t)}{dt} = \sum_k k^\gamma \frac{dN_k(t)}{dt} \, . \tag{4.17c}$$

The first term of (4.17a) accounts for the influx of new nodes of degree 1—attributable to node creation events—, while the second term accounts for the loss of degree 1 nodes—through their acquisition of new edges. The positive term of (4.17b) track this growth in the next compartment, while the negative of (4.17b) represents the loss of nodes—again through their acquisition of edges. The same phenomenology applies for all k, and the system is therefore complete.

The specific form of the rate through compartment k comes from the definition of the growth mechanism. At least one existing node is selected for growth at each time-step, and a second existing node can be selected, with probability $1-b$, yielding a base rate of $[1 + (1-b)]$. To obtain the complete expression of the rate through compartment k, $[1 + (1-b)]$ is multiplied by the fraction of events that affect nodes in this compartment. In a preferential model with kernel k^γ, the probability that a growth event will affect a node of degree k is equal to $N_k k^\gamma / Z_\gamma(t)$, by definition of the model. This yields the final form of (4.17b).

We validate the mean-field equations in Fig. 4.7. We find that the predictions are accurate in most regimes, with the exception of $\gamma \gg 0$, $b \ll 1$, where we find a significant deviation in the tail of the distribution. This discrepancy can be traced back to a strong "peloton dynamics" [60, 98, 248], i.e., a phenomenon whereby a few individuals quickly accumulate a large fraction of the available resources. This effect is notably hard to capture with compartmental mean-field descriptions, and its impact is most felt in the regime $\gamma \gg 0$, $b \ll 1$, where self-loops allow the leader to self sustain.

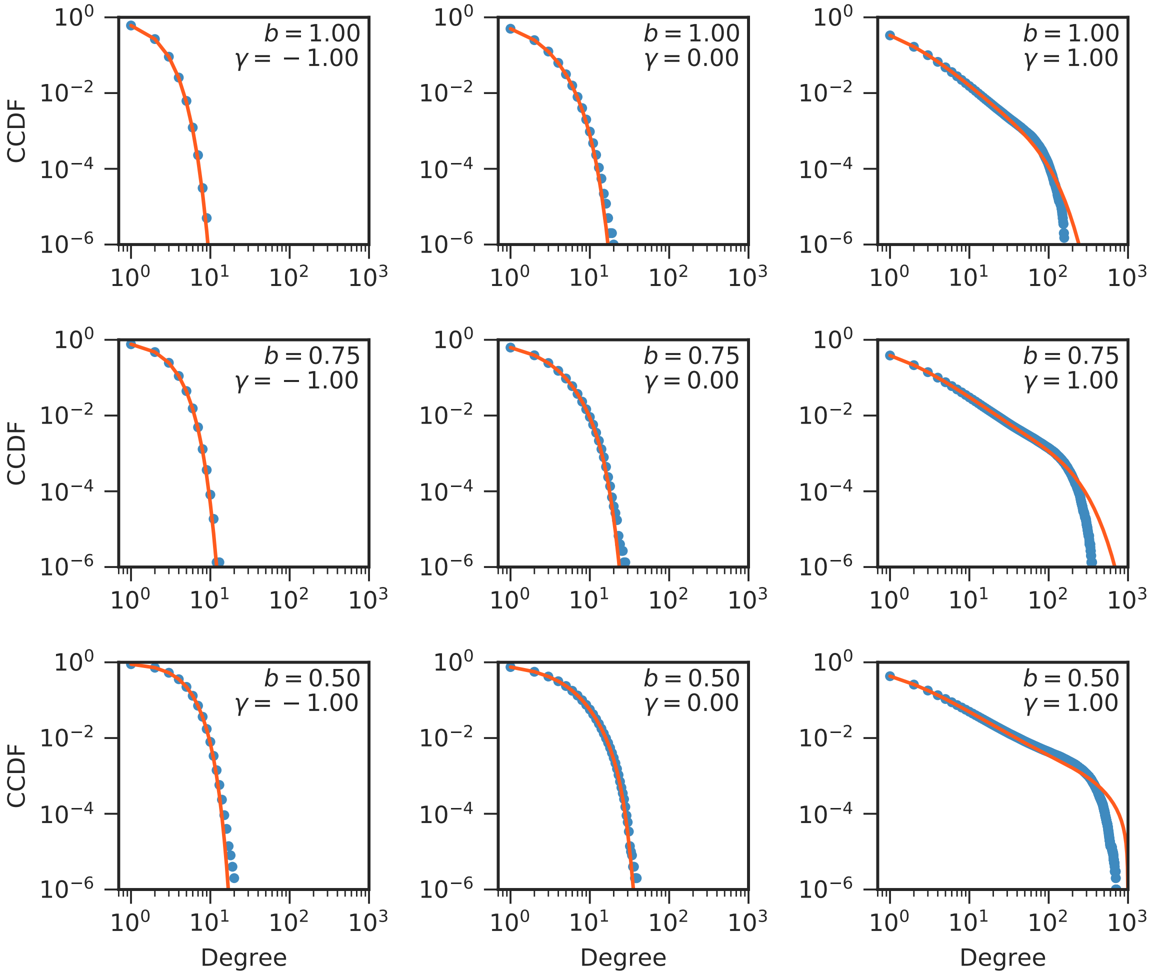


FIGURE 4.7 – Degree distributions of the (γ, b) generalization of PA. Average empirical complementary cumulative distribution functions (blue symbols) versus the solution of Eqs. (4.17) (orange solid lines), for $T = 1000$, $b \in \{1, 0.75, 0.5\}$ (from top to bottom), and $\gamma \in \{-1, 0, +1\}$ (from left to right). Each empirical CCDF is computed on a concatenated degree sequences, obtained from 1 000 instances of the generative model.

Nearly simple graphs in the large network limit

It is clear that the model generates multigraphs with probability bounded away from 0 for all $b < 1$. This is simply due to the fact that new connections between existing nodes are made without regard to the identity of the nodes—doing otherwise by, e.g., rejecting proposed multi-edges and self-loops, would lead to identifiability issues for the parameter and the time-scale [188]. The net effect is that during the first several time-steps, a few densification events will invariably connect nodes to themselves and add redundant connections. But for most choices of exponent γ, this tendency eventually dies out, and the fraction of edges that are redundant edges or self-loops goes to zero with $T \to \infty$ (see Fig. 4.8). Thus, while they are technically multigraphs, the instances of the model are in fact *nearly* simple in the large net-

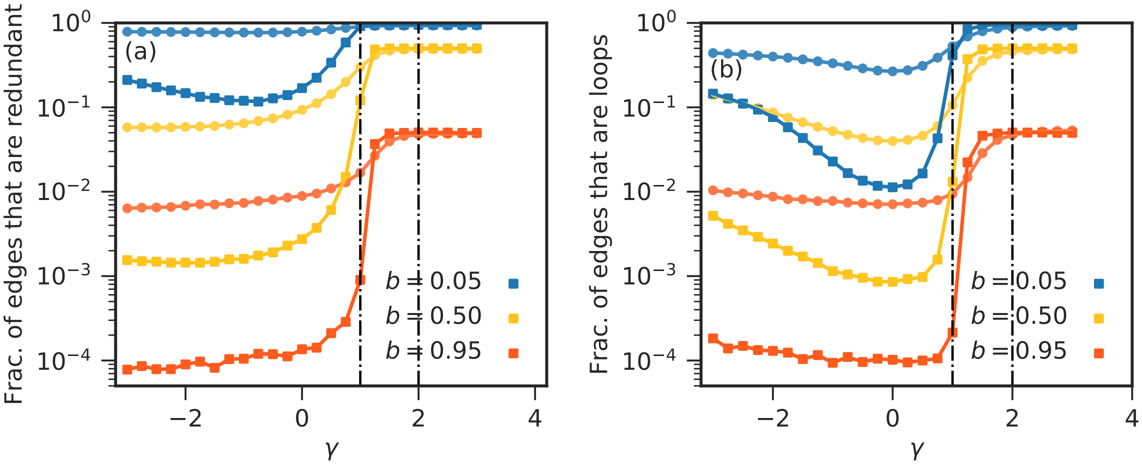


FIGURE 4.8 – Average fraction of the total number of edges that are (a) multiedges, and (b) self-loops, as a function of γ, for three different values of b (denoted by colors), and two different sizes (denoted by marker types). Circles show the fractions for networks of $T = 10^2$ edges, while squares denote networks of $T = 10^4$ edges. In the special case of $b = 1$, it is known that networks start to "condensate" at $\gamma > 1$ (i.e., a finite fraction of the nodes have a macroscopic portion of the edges), and that they fully collapse at $\gamma = 2$ [128]. These two critical values of γ are shown with vertical lines. The curves are obtained by averaging over 1000 instances at each point for $T = 10^2$, and 100 instances at each point for $T = 10^4$.

work limit. Our generalization of PA therefore appears a reasonable model of multigraphs, but also a good approximation of large, sparse networks, with few or no redundant edges and self-loops. This being said, the fraction of redundant edges and self-loops diminishes extremely slowly when b is small ; one should thus reject the (γ, b) generalization of PA as a model if the ratio $|V|/|E| \propto \langle k \rangle^{-1} \ll 1$ and the network is small.

On thresholds. Our numerical results show that the transition to the nearly simple regime occurs somewhere between $\gamma = 1$ and $\gamma = 2$, at a value $\gamma_c(b)$ that depends on b. We do not derive rigorous bounds here, but note that the analysis of Ref. [128] implies the bound $\gamma_c(b) \leq 2$ independent from b. Indeed, we know that the networks *condensate* at $\gamma = 2$ when $b = 1$ (i.e., only a finite fraction of the nodes have degree greater than 1, and we are in a "winner-takes-all" scenario). It is clear that a condensate state also exists at $\gamma = 2$ when $b < 1$, since changing b only gives more opportunities to the leader to gain new connections (with itself, forming self-loops). This condensate will contain an extensive number of loops, which means that the graphs are never simple at $\gamma = 2$, independent from b.

Endogenous correlations

The simple so-called "topological estimator" introduced in the main text rely on the natural correlations that arise between an edge's arrival time and its structural role in the network.

We illustrate these correlations numerically, in Fig. 4.9–4.10. The results shown in Fig. 4.9 are in line with the well-known fact that there are endogenous correlations between the age and the degree of nodes in attachment models (see, e.g., Refs. [4, 16, 127, 128]). Figure 4.10 shows that the onion layer of an edge is even more strongly correlated with its age than the degree of its nodes. This partially explains why the layer-based estimation method performs better than the degree-based method.

We also note that Fig. 4.9–4.10 provide a visual explanation of the behavior of the estimators in the extreme regimes of $\gamma \to -\infty$ and $\gamma \to \infty$. In the regime $\gamma \ll 0$, the degree based estimators perform poorly—but still extract *some* information—because the correlations start to vanish; notice how there are fewer classes the more negative γ becomes. This does not happen with OD; in fact the number of classes appears to grow with increasingly negative values of γ. Thus, the OD-based method can actually better differentiate edges in the regime $\gamma \ll 0$, using the endogenous correlations alone. In contrast, both estimators fail in the regime $\gamma \gg 0$ (see main text). Figures 4.9–4.10 show that this is due to the fact that almost all correlations vanish at the condensation threshold $\gamma = 2$.

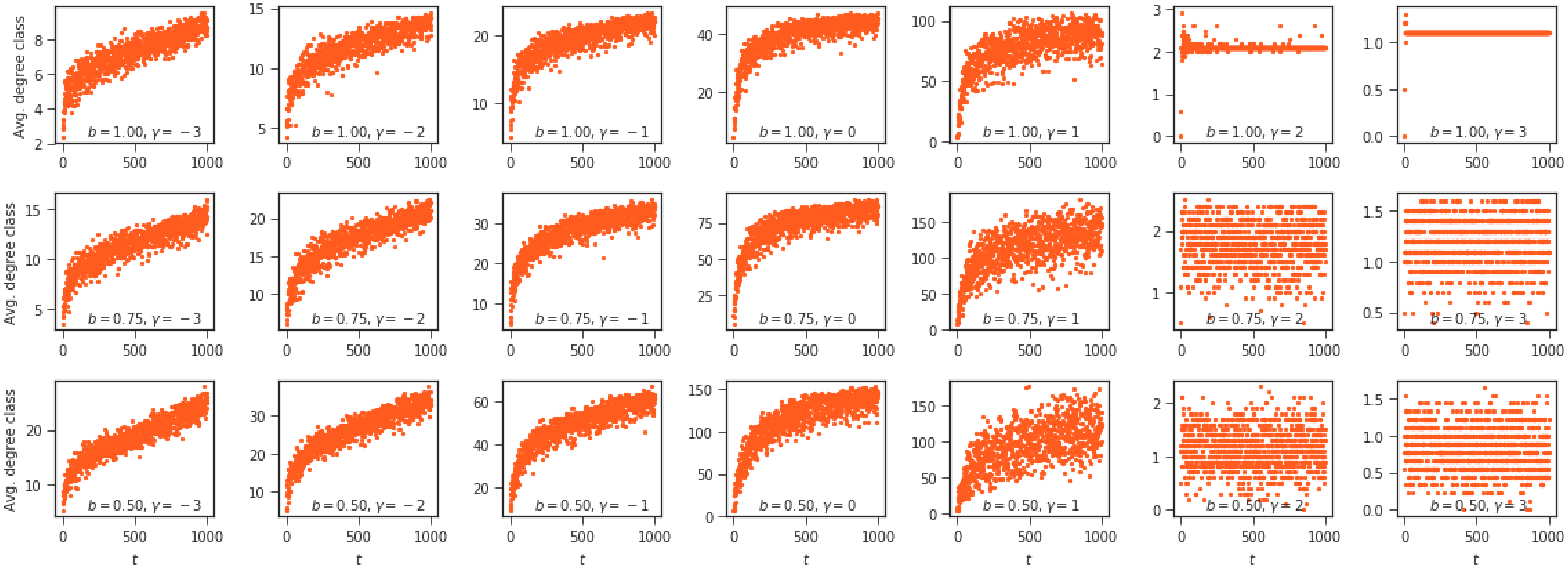


FIGURE 4.9 – True time of arrival of an edge t versus its expected degree class, as determined by the degree sorting algorithm presented in the Materials and Methods section of the main text, for $\gamma = \in \{-3, -2, -1, 0, 1, 2, 3\}$ (from left to right) and $b \in \{0.50, 0.75, 0.95\}$ (from top to bottom). Ten realizations of the growth and inference process are used to compute the expected classes.

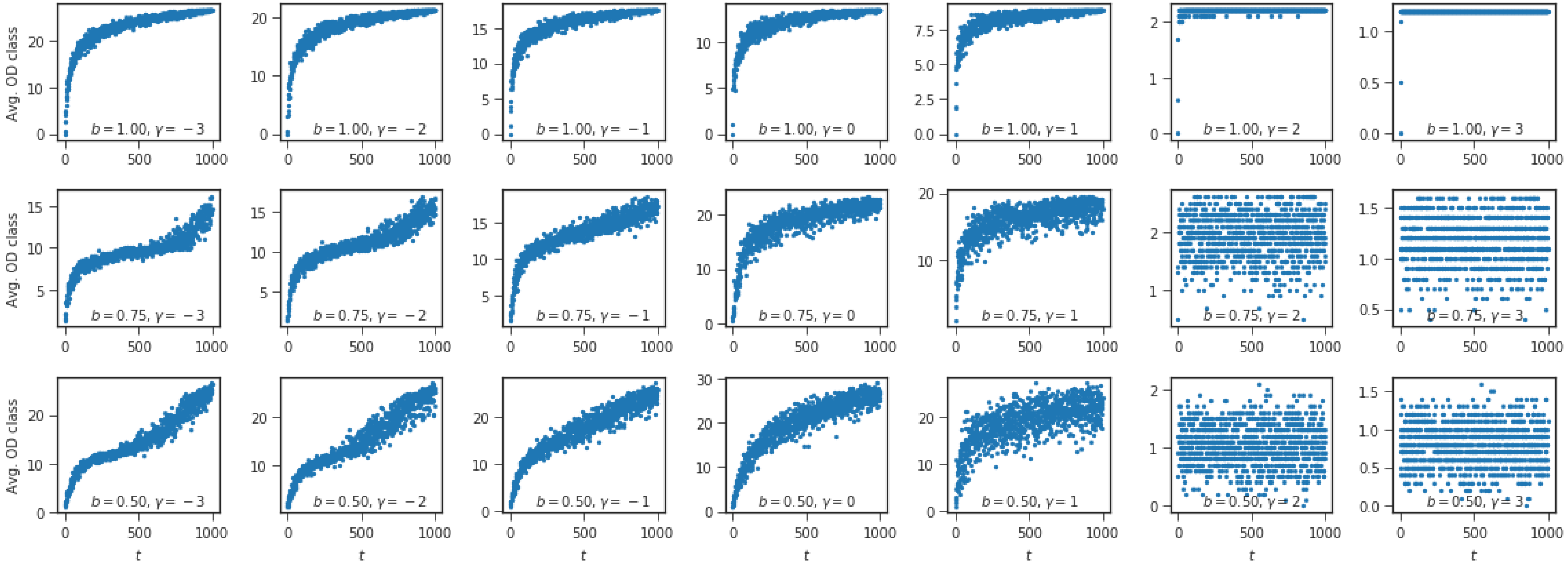


FIGURE 4.10 – True time of arrival of an edge t versus its expected onion decomposition (OD) class, as determined by the OD sorting algorithm presented in the Materials and Methods section of the main text, for $\gamma =\in \{-3,-2,-1,0,1,2,3\}$ (from left to right) and $b \in \{0.50, 0.75, 0.95\}$ (from top to bottom). Ten realizations of the growth and inference process are used to compute the expected classes.

Discriminative power of network properties

The analysis of the previous section suggests that the discriminative power of a method can be imparted to its ability to separate edges in classes that are correlated with age. Thus, having investigated the correlation of the structural properties with an edge's age, we turn to the separating power of these properties.

Given a property that partitions the edge set E in q disjoint classes $C_1, C_2, \ldots C_k$ (i.e. classes that satisfy $C_i \cap C_j = \emptyset$ for all $i \neq j$, but also $\bigcup_{i=1}^{q} C_i = E$), we define the partition entropy as

$$S = -\sum_{i=1}^{q} \frac{|C_i|}{|E|} \log_2 \frac{|C_i|}{|E|} = \log_2 T - \frac{1}{T}\sum_{i=1}^{q} |C_i| \log_2 |C_i| \,. \tag{4.18}$$

This entropy S characterizes the information content (in bits) of the partition induced by the graph property : The larger S, the more discriminative it is. A hypothetical ideal property would be associated with an entropy of $S = \log_2 T$ (q classes of size one). Any property that operates with a smaller information content cannot resolve all edges.

We show in Fig. 4.11 the entropy of the partitions for each property, as a function of γ, for many values of b. Focusing first on the regime $b = 1$, we see that the entropy is not a perfect predictor of the "quality" of a property, if it is analyzed in isolation. For example, the entropy of the degree partition is extremely large at $\gamma = 1$—much larger than that of the onion partition. Based on this observation, we could be tempted to conclude that OD will perform poorly in comparison with the degree. But this turns out not to be the case : We know that OD yields a better estimate of the arrival than the degree at that value of γ (see main text). This apparent contradiction is due to the fact that even if the degree is more discriminative than the onion decomposition, its induced ranking correlates poorly with the age, while OD's does not (see Figs. 4.9–4.10). That is not to say, however, the entropy is useless. We can understand the good performances of OD in the regime $\gamma \ll 0$ in terms of entropy : The entropy grows without apparent bounds in this regime. Therefore, the OD partition keeps on gaining discriminative power, which the correlation plots of Fig. 4.10 shows to be 'good' discriminative power (i.e., correlated with age).

In the regime $b < 1$, Fig. 4.11 shows that the discriminative power of the OD method is not nearly as large ; that is to say, the entropy of the onion partition does not grow without bound with increasingly negative values of γ. This is in line with the results of the main text that show smaller achievable correlation (by OD), when the value of b is small and $\gamma < 0$. In contrast, the entropy of the degree partition is hardly affected ; This observation, in conjunction with Fig.4.9, explains the relatively good performances of the degree estimator in the regime $b < 1$.

Finally, we note that the entropy of the two partitions goes to some minimum constant at $\gamma = 2$, for all b. This is again due to the apparition of a condensate state. When $b = 1$, the

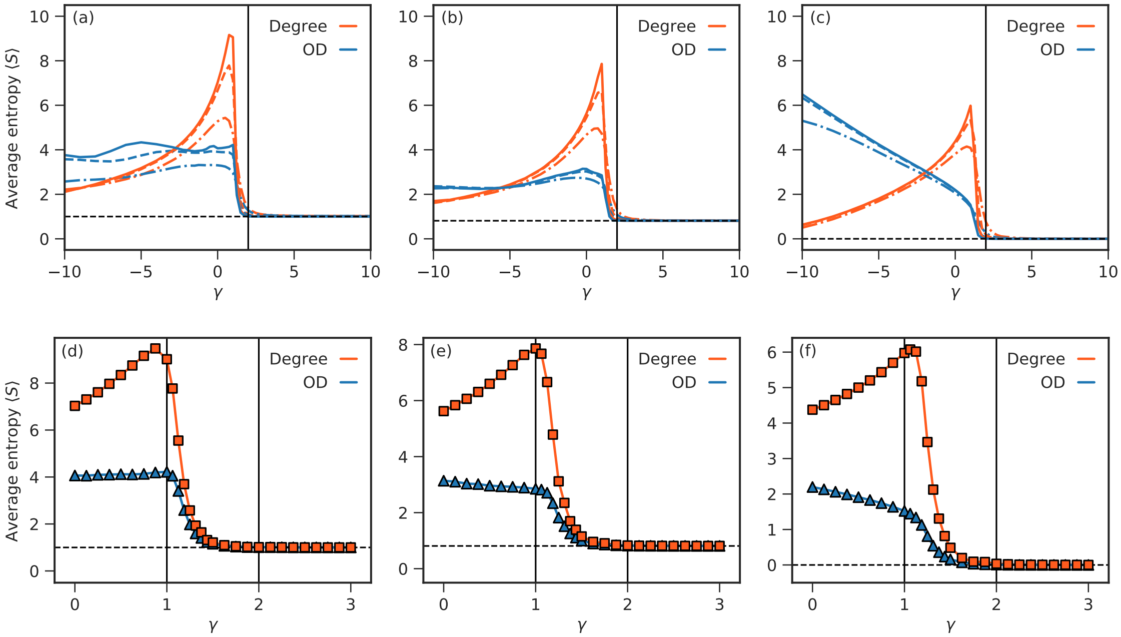


FIGURE 4.11 – Average entropy of the edge partitions as a function of the model parameters. The entropy S characterizes the information content of the partition induced by a property : The larger $\langle S \rangle$ at a point (γ, b), the more discriminative the structural property at that point. (a-c) Average partition entropy of the degree classes (orange) and the OD classes (blue), at (a) $b = 0.5$, (b) $b = 0.75$, and (c) $b = 1.00$. Solid lines are obtained with graphs of $T = 10^4$ edges, while dotted lines correspond to smaller instances (of $T = 10^2$ and $T = 10^3$ edges). We show the upper bound on the condensation threshold with a solid vertical line at $\gamma = 2$, and the entropy $h(b) = -b \log b - (1-b)\log(1-b)$ of the condensate with a dotted horizontal line. (d-f) Details of figures (a-c) in the interval $\gamma \in [0,3]$, computed at $T = 10^5$, with $b = 0.5$ (d), $b = 0.75$ (e), and $b = 1.00$ (f). This interval contains many qualitative phase transitions [128], two of which are shown using vertical lines (the scale-free regime at $\gamma = 1$ and the upper bound on the condensate at $\gamma = 2$). Again, the entropy of the condensate is indicated with a horizontal line. In all figures, the curves are obtained by averaging over 1000 realizations of the process when $T = 10^2$ and $T = 10^3$, and 100 realizations when $T = 10^4$.

condensate is roughly a star-graph, and there is therefore only one edge class, which implies a null entropy. When $b < 1$, the condensate is a star-graph, augmented by many self-loops. For this particular graph, both the degree and onion decomposition based classifiers put the self-loops in one class, and the rest of the edges in another class. Since we can approximate the fraction of edges that are self-loop with $1 - b$ in the condensate, the entropy should tend to the binary entropy $h(b) = -b \log_2 b - (1-b)\log_2(1-b)$ in the regime $\gamma > 2$. This is confirmed in Fig. 4.11.

4.8.2 On uniform posterior distributions

In the main text, we mention that the posterior distribution $P(X|G,\gamma,b)$ is often uniform over large sets of histories, and that this epitomizes why likelihood maximization is not a suitable method to extract information from G. We will now demonstrate that the Krapivsky-Redner-Leyvraz generalization of preferential attachment [128] is an extreme example of this problem. Namely, we will explicitly show that when $\gamma \in \{0,1\}$, the posterior distribution of this model is uniform over all histories in $\Psi(G)$ [1]. We will then further argue that there exists large equivalence classes for general $\gamma \in \mathbb{R}$ and $b \in [0,1]$.

Recall from the Materials and Methods that in the model where $b = 1$, the logarithm of the prior distribution is given by

$$\log P(X|\gamma) = \sum_{t=1}^{T-1} \log w_{a_t}(\gamma, G_t) \, , \qquad w_i(\gamma, G_t) = \frac{k_i^\gamma(t)}{\sum_{i\in V_t} k_i^\gamma(t)} \, , \tag{4.19}$$

where $G_0, \ldots, G_{T-1}$ is the sequence of graphs described by history X, V_t is the node set of G_t prior to any modification of the graph's structure, and where we have simplified the notation by writing a_t to denote the node selected as the fixation site at time t in X. To demonstrate that the distribution is uniform over the set of all consistent histories, we first define the normalization $Z_\gamma(t) = \sum_{i\in V_t} k_i^\gamma(t)$ and rewrite the prior as

$$\log P(X|\gamma) = \sum_{t=1}^{T-1} \left[\log k_{a_t}^\gamma(t) - \log Z_\gamma(t, G_t)\right] \, . \tag{4.20}$$

Now, in the special cases of uniform attachment and preferential attachment, corresponding to $\gamma = 0$ and $\gamma = 1$, the normalization $Z_\gamma(t)$ *always* takes a special value independent from the actual content of V_t, namely :

$$Z_0(t) = \sum_{i\in V_t} k_i^0 = |V_t| = t + 1 \, , \tag{4.21a}$$

$$Z_1(t) = \sum_{i\in V_t} k_i = 2t \, . \tag{4.21b}$$

The second identity follows from the fact that exactly one edge is created at each t, and that the sum of all degrees is always equal to twice the number of edges. These normalizations are independent of X, meaning that they can be dropped as an additive constant, identical for all X. Then, using $\gamma = 0$ and $\gamma = 1$ in Eq. (4.20), we are left with

$$\log P(X|\gamma) \propto \begin{cases} T-1 & \gamma = 0, \\ \sum_{t=1}^{T-1} \log k_{a_t}(t) & \gamma = 1. \end{cases} \tag{4.22}$$

This last equation directly shows that the prior distribution (and therefore the posterior distribution) is uniform over all histories when $\gamma = 0$. Less obvious is the fact that the equation also implies a uniform posterior distribution in the case $\gamma = 1$. To see this, notice that

[1] Lemma 5.3 of Ref. [147] states this fact without proof.

the posterior distribution is obtained by conditioning on G, and that this restricts the possible histories to those in which a node i of degree k_i^* in G appears $k_i^* - 1$ times in the sum $\sum_{t=1}^{T-1} \log k_{a_t}(t)$: Once as a node of degree one, once as a node of degree two, etc. Thus, each history consistent with G will be associated with some permutation of the sum. Obviously, a permutation does not change the value of a sum ; therefore, the posterior distribution is uniform over all histories consistent with G.

Large sets of equally likely histories also arise in the more general attachment model on trees (i.e., when $\gamma \in \mathbb{R}$ with $b = 1$). This again forbids the use of likelihood maximization as an inference technique. The proof that these sets of histories exist is similar in spirit to that of the special cases above. We first make use of the permutation argument again, noting that it applies to the general sum $\sum_{t=1}^{T-1} \log k_{a_t}^{\gamma}(t)$, regardless of the value of γ. The problem therefore reduces to the study of the evolution of the normalization constant. Different from the special cases $\gamma = 0$ and $\gamma = 1$, the normalization $Z_\gamma(t)$ does not grow at the same rate for all histories when γ is arbitrary. But, as we now show, this does not preclude the existence of equivalence classes with respect to the posterior distribution. For example, consider two histories identical in all respects until a last node of degree k and its $k - 1$ remaining neighbors are encountered. The $(k - 1)!$ histories resulting from the enumeration of this neighborhood will have, by construction, equivalent sequences of normalization constants $Z_\gamma(1) \to Z_\gamma(2) \to \ldots Z_\gamma(T - 1)$, which imply that these histories will be associated with the same posterior probability, and that they will form a small equivalence class. This is a somewhat contrived example, but there are many more possibilities. For instance, one could apply the same argument at any point of any history (not only at the end), or *combine* the enumeration of m different neighborhoods to yield $k_1! \times k_2! \times \ldots \times k_m!$ different histories with the same posterior probability. Therefore, it follows that the posterior distribution of the model $(\gamma \in \mathbb{R}, b = 1)$ is uniform over large set of histories, as was claimed. A nearly identical derivation, here omitted, provides a proof that this is also the case for general $b \in [0,1]$.

Let us note in closing that the uniformity results of the models $\gamma \in \{0,1\}$ with $b = 1$ can be seen as arising from the relationship between random recursive trees and growth processes, see Ref. [64, pp. 14–16]. Specifically, it is well-known that generating a graph of T edges with $(\gamma = 0, b = 1)$ is equivalent to drawing a graph uniformly from the set of all non-plane recursive trees (of T edges), while generating a graph of T edges with $(\gamma = 1, b = 1)$ is equivalent to drawing a graph uniformly from all plane recursive trees (of T edges). This directly implies the uniformity of the posterior distribution ; since the prior $P(X|\gamma)$ is itself uniform over the larger set of all trees. The conditioning on G via the likelihood $P(G|X, \gamma)$ merely renormalizes the distribution to the set of histories consistent with G, preserving uniformity.

4.8.3 Properties of the average posterior time of arrival

Minimum mean-square error

In the main text, we mention without proof that the posterior average of $\tau_X(e)$ minimizes the mean-square error on $\tau_{\tilde{X}}(e)$. The proof is standard and goes as follows.

We first assume that the true arrival time of edge e is determined by its posterior distribution $p_e(t|G,\gamma,b)$. Without access to the ground-truth, this is the best guess we can make since the posterior distribution extracts all available information from the graph G. The mean-square error (MSE) associated with some estimator $\hat{\tau}(e)$ of the true arrival time $\tau_{\tilde{X}}(e)$ of edge e is then

$$\begin{aligned}\mathrm{MSE} &= \int dt(\hat{\tau}(e)-t)^2 p_e(t|G,\gamma,b)\\ &= \sum_{X\in\Psi(G)} P(X|G,\gamma,b)\int dt\big(\hat{\tau}(e)-t\big)^2\delta(t-\tau_X(e))\\ &= \sum_{X\in\Psi(G)} P(X|G,\gamma,b)\big(\hat{\tau}(e)-\tau_X(e)\big)^2\,,\end{aligned}$$

where we have used the definition $p_e(t|G,\gamma,b)=\sum_{X\in\Psi(G)}P(X|G,\gamma,b)\mathbb{I}[\tau_X(e)=t]$, and where we have written the indicator function as Dirac's delta.

Even if the timescale of the process is by definition discrete, we resort to continuous estimators of $\tau_{\tilde{X}}(e)$. This is motivated not only by the simplifications that this decision brings to the mathematical treatment of the problem, but also by the fact that the generated graphs almost never encode the temporal ordering perfectly [146]. In short, total ordering won't do [147], and continuous estimators are needed to encode our imperfect guesses.

With this in mind, the estimator $\hat{\tau}^{\mathrm{MMSE}}(e)$ that minimizes the mean-square error must solve

$$\frac{\partial}{\partial\hat{\tau}(e)}\left[\sum_{X\in\Psi(G)} P(X|G,\gamma,b)\big(\hat{\tau}(e)-\tau_X(e)\big)^2\right]_{\hat{\tau}(e)=\hat{\tau}^{\mathrm{MMSE}}(e)}=0\,, \tag{4.23}$$

or more explicitly

$$\sum_{X\in\Psi(G)} P(X|G,\gamma,b)\big(\hat{\tau}^{\mathrm{MMSE}}(e)-\tau_X(e)\big)=0\,. \tag{4.24}$$

Using the normalization of the posterior distribution, one easily obtains

$$\hat{\tau}^{\mathrm{MMSE}}(e)=\langle\tau_X(e)\rangle\,, \tag{4.25}$$

where the average is taken over the posterior distribution for X.

Maximal correlation

A similar line of reasoning can be used to show that the MMSE estimators maximize the correlation, on average. By a slight abuse of notation, let us refer to an estimated history constructed with $\{\hat{\tau}(e)\}_{e\in E(G)}$ as Y, such that $\tau_Y(e) = \hat{\tau}(e)$. Then, assuming again that X is drawn from the posterior distribution, the expected correlation of X and Y is

$$\begin{aligned}\langle\rho(X,Y)\rangle &= \sum_{X\in\Psi(G)} P(X|G,\gamma,b)\rho(X,Y)\\ &= \frac{1}{\sigma_X\sigma_Y}\sum_{X\in\Psi(G)} P(X|G,\gamma,b)\sum_{e\in E(G)}\big(\tau_X(e)-\langle\tau\rangle\big)\big(\tau_Y(e)-\langle\tau\rangle\big)\\ &= \frac{1}{\sigma_X\sigma_Y}\sum_{e\in E(G)}\big(\langle\tau_X(e)\rangle-\langle\tau\rangle\big)\big(\tau_Y(e)-\langle\tau\rangle\big)\,, \end{aligned}\tag{4.26}$$

where we have defined $\sigma_X^2 := \sum_{e\in E(G)}(\tau_X(e)-\langle\tau\rangle)^2$, and σ_Y similarly. These standard deviations can be taken out of the sum because σ_Y is independent of X and the value of σ_X is constant for all X. The uniformity of σ_X comes from the fact that one edge must occupy each 'time slot' $t=0,\ldots,T-1$ by definition, which implies $\sigma_X^2 = \sum_{t=0}^{T-1}(t-\langle\tau\rangle)^2$. Again, somewhat stretching the notation, we define as Z the history constructed with the MMSE estimators, i.e, the history such that $\tau_Z(e) = \langle\tau_X(e)\rangle$. This allows us to express the expected correlation compactly as

$$\langle\rho(X,Y)\rangle = \frac{\sigma_Z}{\sigma_X}\rho(Z,Y)\,,\tag{4.27}$$

where $\sigma_Z \neq \sigma_X$ in general. In other words, we find that the expected correlation is proportional to the correlation of our arbitrary estimators $\{\tau_Y(e)\}$ and the MMSE estimators $\{\langle\tau_X(e)\rangle\}$. This implies that the expected overall correlation will be maximized by using the MMSE estimators for the arrival times.

Equation (4.27) has additional interesting consequences. First, it gives a compact expression of the expected correlation achieved by the MMSE estimators, since using the MMSE estimators amounts to setting $Z=Y$, yielding

$$\langle\rho(X,Z)\rangle = \frac{\sigma_Z}{\sigma_X}\,.\tag{4.28}$$

Second, (4.27) confirms the intuition that a good correlation can only be achieved by reliably ordering all the edges. Indeed, for the MMSE estimators the ratio σ_Z/σ_X gives an expected correlation, i.e., an average of numbers in $[-1,1]$. It follows that this ratio can never be greater than 1. This implies $\sigma_Z \leq \sigma_X$, where σ_X^2 is a known variance (see above). Now, turning this statement around : The ratio σ_Z/σ_X will be maximized if the (MMSE) estimator achieves a variance σ_Z^2 equal to the variance σ_X^2, obtained by ordering all edges. Grouping edges in equivalence classes reduces the variance σ_Z (since grouping two MMSE estimators imply averaging their ranking); this implies a degradation of the correlation, since we move away from σ_X in so doing. Thus, the maximum correlation $\langle\rho(X,Z)\rangle = 1$ can only be achieved if the MMSE estimators give a total ordering of the edges.

4.8.4 Properties of the correlation

Impossibility of improving on equivalence classes

Suppose that some estimator of $\tau_{\tilde{X}}(e)$ cannot differentiate the edges in a subset $S \subset E$, and that according to the estimation procedure, these edges would occupy time slots $t+1$ through $t+m$ (where $m = |S|$). We now show that trying to arbitrarily break the tie will not improve correlation on average.

We define the history Y as the one where $\tau_Y(e) = t + (m+1)/2$ if $e \in S$. We also consider the $m!$ different ways of randomly breaking the ties for S (i.e., the permutations of S). We denote these histories with $\{Z_{\pi_i}(X)\}$. They satisfy the property that $\tau_{Z_{\pi_i}}(e) = \tau_Y(e)$ if $e \notin S$, and that the edges of S are assigned some permutations of the times of arrival $t+1, \ldots, t+m$, indexed by i.

If we break the ties randomly, we will achieve, on average, a correlation of

$$\langle \rho(X, Z_\pi) \rangle_\pi := \frac{1}{m!} \sum_{i \in \pi(S)} \rho(X, Z_{\pi_i}) \tag{4.29}$$

with the ground truth X (here noted without tilde to simplify the notation). Then, to prove that randomly breaking ties cannot improve the correlation, we must show that

$$\langle \rho(X, Z_\pi) \rangle_\pi - \rho(X, Y) := \Delta \leq 0\,. \tag{4.30}$$

By symmetry of the permutation group, we can compute the average as

$$\langle \rho(X, Z_\pi) \rangle_\pi = \frac{1}{2} [\rho(X, Z_{\pi_+}) + \rho(X, Z_{\pi_-})]\,, \tag{4.31}$$

where Z_{π_-} is the history constructed with some permutation of S and where Z_{π_+} contains the inverse permutation. Noting that $\sigma_{Z_{\pi_-}} = \sigma_{Z_{\pi_+}}$, we write $\sigma_X \sigma_{Z_{\pi_+}} \Delta$ as

$$\begin{aligned} \frac{1}{2} \left[\sum_{e \in E(G)} \big(\tau_{Z_{\pi_-}}(e) - \langle \tau \rangle\big)\big(\tau_X(e) - \langle \tau \rangle\big) + \sum_{e \in E(G)} \big(\tau_{Z_{\pi_+}}(e) - \langle \tau \rangle\big)\big(\tau_X(e) - \langle \tau \rangle\big) \right] \\ - \frac{\sigma_{Z_{\pi_+}}}{\sigma_Y} \left[\sum_{e \in E(G)} \big(\tau_Y(e) - \langle \tau \rangle\big)\big(\tau_X(e) - \langle \tau \rangle\big) \right]. \end{aligned} \tag{4.32}$$

The standard deviations σ_X and $\sigma_{Z_{\pi_+}}$ are non-negative coefficients by definition. It therefore suffices to show that the above expression is smaller than or equal to 0.

To do so, we first note that $\sigma_{Z_{\pi_+}} > \sigma_Y$[2]. This implies that

$$\sigma_X \sigma_{Z_{\pi_+}} \Delta < \sum_{e \in S} \Big[\frac{1}{2} \big(\tau_{Z_{\pi_-}}(e) - \langle \tau \rangle \big) \big(\tau_X(e) - \langle \tau \rangle \big) \\ + \frac{1}{2} \big(\tau_{Z_{\pi_+}}(e) - \langle \tau \rangle \big) \big(\tau_X(e) - \langle \tau \rangle \big) - \big(\tau_Y(e) - \langle \tau \rangle \big) \big(\tau_X(e) - \langle \tau \rangle \big) \Big] \quad (4.33)$$

where we have used $\sigma_{Z_{\pi_+}} / \sigma_Y > 1$ and the fact that the permuted histories are identical to Y for all $e \notin S$. To obtain an explicit numerical expression for the right-hand side, we index the edges of S by their true time of arrival (i.e., their arrival time in X). That is to say, we choose

$$\tau_X(e_1) < \tau_X(e_2) < \ldots < \tau_X(e_m) \, . \quad (4.34)$$

With an ordering defined, we choose the following permutations to construct Z_{π_-} and Z_{π_+} :

$$\begin{array}{ccccc} Z_{\pi_+} : & t+1 & t+2 & \ldots & t+m \\ & \downarrow & \downarrow & \cdots & \downarrow \\ & e_1 & e_2 & \cdots & e_m \end{array}$$

$$\begin{array}{ccccc} Z_{\pi_-} : & t+1 & t+2 & \ldots & t+m \\ & \downarrow & \downarrow & \cdots & \downarrow \\ & e_m & e_{m-1} & \cdots & e_1 \end{array}$$

In other words, in Z_{π_+}, edge e_1 is assigned time $t+1$, edge e_2 is assigned time $t+2$, etc. Then, recalling that $\tau_Y(e) = t + (m+1)/2$ for all $e \in S$, we obtain for the right-hand side of (4.33) :

$$\sum_{i=1}^{m} \Big[\frac{1}{2} \big(t + m - i + 1 - \langle \tau \rangle \big) \big(\tau_X(e_i) - \langle \tau \rangle \big) \\ + \frac{1}{2} \big(t + i - \langle \tau \rangle \big) \big(\tau_X(e_i) - \langle \tau \rangle \big) - \big(t + (m+1)/2 - \langle \tau \rangle \big) \big(\tau_X(e_i) - \langle \tau \rangle \big) \Big] \\ = \sum_{i=1}^{m} \Big[\frac{1}{2} \big(t + m - i + 1 - \langle \tau \rangle \big) + \frac{1}{2} \big(t + i - \langle \tau \rangle \big) - \big(t + (m+1)/2 - \langle \tau \rangle \big) \Big] \big(\tau_X(e_i) - \langle \tau \rangle \big) = 0$$

This completes the proof that randomly breaking ties will on average yield a worst outcome than merely guessing $\tau_Y(e) = t + (m+1)/2$ for the edges $e \in S$.

On the placement of equivalence classes

Consider again a set of tied edges S with prescribed arrival times $t+1, ..., t+m$ (if they were not tied). We have shown in the above section that assigning $\tau_Y(e) = t + (m+1)/2$ for all

[2] This can also be shown directly by comparing the *variances*; one finds that $\sigma^2_{Z_{\pi_+}} - \sigma^2_Y = \sum_{e \in S} [\tau^2_{Z_{\pi_+}}(e) - \tau^2_Y(e)] > 0$, since $\tau_Y(e) = t + (m-1)/2$ for all $e \in S$ and the $\tau_{Z_{\pi_+}}$ goes from $t+1$ to $t+m$.

$e \in S$ is a better choice, on average, than breaking the ties at random. We now show this choice is in fact a good rule of thumb, if the optimal MMSE estimators are unknown.

For the purpose of this demonstration, we assign the time $\lambda + \varepsilon$ to the edges in S, where $\lambda = t + (m+1)/2$ and where ε is a small perturbation. Assuming that all edges *not in S* are assigned a unique time, or that the other equivalence classes have been collapsed on *their* averages, the overall average time of arrival of the edges in history Y is

$$\langle \tau(\varepsilon) \rangle = \frac{1}{T}\left[\sum_{i=0}^{t} i + m(\lambda+\varepsilon) + \sum_{i=t+m+1}^{T-1} i\right] = \frac{m}{T}(\lambda+\varepsilon) + C =: \langle\tilde{\tau}\rangle + \frac{m}{T}\varepsilon \tag{4.35}$$

where $\langle\tilde{\tau}\rangle$ would be the average if $\varepsilon = 0$. Now, the variance of the arrival times in $Y(\varepsilon)$ is

$$\sigma_Y^2(\varepsilon) = \sum_{e\in E}\left(\tau_Y(e) - \langle\tilde{\tau}\rangle - \frac{m\varepsilon}{T}\right)^2 = \sigma_Y^2(0) + 2m\varepsilon(\lambda - \langle\tilde{\tau}\rangle) + m\epsilon^2\left(1 + \frac{m}{T}\right) , \tag{4.36}$$

such that the standard deviation is, to the first order in ε,

$$\sigma_Y(\varepsilon) = \sigma_Y(0)\left[1 + \frac{m\varepsilon(\lambda - \langle\tilde{\tau}\rangle)}{\sigma_Y^2(0)} + O(\varepsilon^2)\right] . \tag{4.37}$$

This variance can be substituted in Eq. (4.27) to compute the expected correlation of the history $Y(\varepsilon)$ with the ground-truth X, as a function of ε.

Noting that $\tau_Y(e) = \lambda + \varepsilon$ if $e \in S$, and that the sum of $\tau_Z(e) - \langle\tilde{\tau}\rangle$ over all edges is equal to 0, we find

$$\begin{aligned}
\langle \rho(X,Y)\rangle &\propto \frac{\sum_{e\in E}(\tau_Z(e) - \langle\tilde{\tau}\rangle)(\tau_Y(e) - \langle\tilde{\tau}\rangle - m\varepsilon/T)}{\sigma_Y(\varepsilon)\sigma_Z} \\
&= \left[\langle\tilde{\rho}\rangle + \varepsilon\frac{\sum_{e\in S}(\tau_Z(e) - \langle\tilde{\tau}\rangle)}{\sigma_Y(0)\sigma_Z}\right]\left[1 - \frac{m\varepsilon(\lambda - \langle\tilde{\tau}\rangle)}{\sigma_Y(0)^2} + O(\varepsilon^2)\right] \\
&= \langle\tilde{\rho}\rangle\left\{1 + \varepsilon\left[\frac{1}{\langle\tilde{\rho}\rangle}\frac{\sum_{e\in S}(\tau_Z(e) - \langle\tilde{\tau}\rangle)}{\sigma_Y(0)\sigma_Z} - \frac{m(\lambda - \langle\tilde{\tau}\rangle)}{\sigma_Y(0)^2}\right] + O(\varepsilon^2)\right\}
\end{aligned} \tag{4.38}$$

where $\langle\tilde{\rho}\rangle$ is the averaged correlation achieved when $\varepsilon = 0$. Since in the present scenario, we are explicitly using suboptimal estimators (i.e., not the MMSE, which do not allow for arbitrary choices of λ and ε), we do not have access to $\tau_Z(e)$ for $e \in S$. This implies that it is impossible to know whether the coefficient of ε is positive or negative; a priori the two terms appearing in this coefficient have about the same order of magnitude (recall that $|S| = m$, and that both $\sigma_Y(0)$ and σ_Z are variances with exactly the same number of terms). Thus, the safest bet is to not alter the average, and to choose $\varepsilon = 0$.

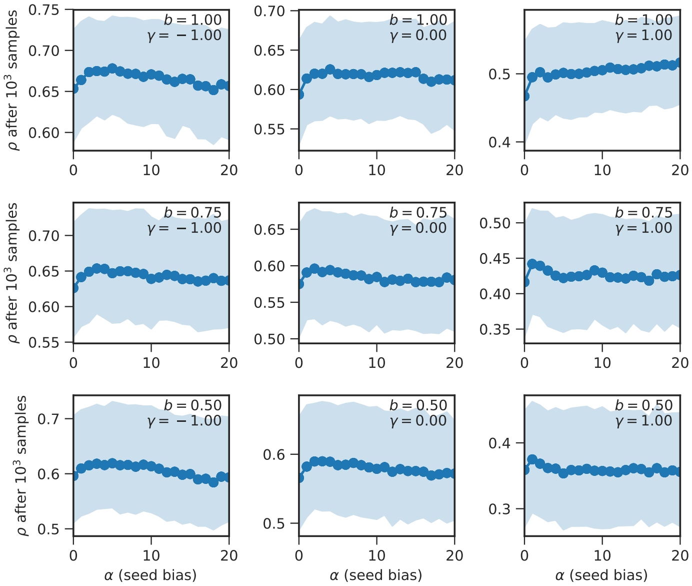


FIGURE 4.12 – Average correlation attained by the unbiased snowball sampler ($\beta = \mu = 0$), with only 10^3 samples, as a function of the strength of the bias α used in selecting the seed. The experiments are run on artificial graphs of $T = 50$ edges generated by the model, with $b \in \{1, 0.75, 0.5\}$ (from top to bottom), and $\gamma \in \{-1, 0, +1\}$ (from left to right). The curves are averaged for 10 runs of the sampler, on 200 different instances of the model (i.e., all averages are computed with 2 000 data points). The shaded region contains 50 percent of the data points (from the 25th percentile to the 75th percentile).

4.8.5 Characterization of the snowball sampling method

A snowball sample is a random recursive enumeration of a graph, rooted on a randomly selected seed [70, 93, 135]. In the main text, we use these samples to compute posterior average for the model, via the importance sampling method ; see the Material and Methods for details. An important caveat of the importance sampling algorithm is that while it works with *any* sampler that can generate all graphs in $\Psi(G)$ with non-zero probability, the algorithm converges faster when the samples approximate the ground-truth more closely. In this section of the Supplementary Material, we therefore examine possible variants of the snowball sampling algorithm, with the goal of generating samples close to the ground-truth.

Effect of the initial bias

The probability of generating a history X with the unbiased snowball sampling algorithm is

$$Q_{\mathrm{sb}}(X|G) = P(e_0|G;\alpha) \times \prod_{t=1}^{T-1} \left[|\Omega_X(t)| \right]^{-1}, \tag{4.39}$$

where $\Omega_X(t)$ denotes the boundary at step t of history X, and where $P(e_0|G)$ is the distribution used to select the seed (the initial edge). The distribution $P(e_0|G)$ is our first and simplest optimization opportunity ; it allows us to start on the right foot, with edges that are likely to be the first. We use the parametrized distribution

$$P(e|G;\alpha) = \frac{(k_1(e) \times k_2(e))^\alpha}{Z(\alpha)} \qquad\qquad Z(\alpha) = \sum_{e \in E(G)} (k_1(e) \times k_2(e))^\alpha \tag{4.40}$$

to fine-tune the strength of this bias [$k_1(e)$ and $k_2(e)$ are the degrees of the nodes attached to e, and α is a bias parameter]. Large values of α nudge the algorithm towards histories that begin with edges that are attached to high degree nodes. Thus the distribution (4.40) makes use of the endogenous correlations highlighted in Fig. 4.9 to better guide the sampler. We investigate the effects of α in Fig. 4.12.

On the small instances ($T = 50$) used in the numerical study of Fig. 4.12, non-zero values of α appear to enhance correlation systematically, if only so slightly. It is in general dangerous to bias a sampler too strongly, since we then run the risk of not properly sampling the relevant space— a generic problem of importance sampling methods [11]. Therefore, as a rule of thumb, and based on the result of Fig. 4.12, we opt for $\alpha = 5$ in all simulations. This value strikes a balance between increased correlation and complete coverage of the space ; it is certainly better than no bias at all. As Fig. 4.12 shows, however, the results may vary depending on the specifics of the parameters. It may be worth characterizing the sampler summarily given a specific graph.

Controlled evolution

Both b and γ are constants of the generative process. This implies that unbiased estimators $(\hat{b}(t), \hat{\gamma}(t))$ constructed with the first t iterations of the ground-truth $\tilde{X}$ ought to settle on (b, γ), for sufficiently large t. In this short section, we investigate a *biased* snowball sampler that makes use of this property to generate better histories.

Our goal in defining this *biased snowball sampler* is to actively control the evolution of $(\hat{b}(t), \hat{\gamma}(t))$. Our reasoning is that if the estimators fluctuate a lot throughout a history X, then X is probably not the ground-truth (per the above observation). In fact, X is unlikely to be even well correlated with the ground-truth. Ergo, we should steer the sampler away from such histories.

We can achieve this kind of control locally, by weighting the edges of the boundary. Indeed, suppose that we have running estimators of b and γ, and that we can compute how each edge of the boundary would change these estimators, if it were selected as the next one. Then we can use these estimators to give more weight (i.e., a larger probability) to edges that will bring the estimators towards their constant baseline. If the control is strong enough, the histories should be associated with nearly constant estimators trajectories.

In practice, running estimators of γ are hard to evaluate, let alone update efficiently (see

Materials and Methods of the main text). Therefore we use the estimator pair :

$$\hat{b}(t) = \frac{|V_t| - 2}{|E_t| - 1} \qquad \text{and} \qquad \hat{v}(t) = \frac{1}{|V_t|}\left[\sum_{i\in V_t} k_i(t)\right]^2 + \frac{1}{|V_t|^2}\sum_{i\in V_t} k_i^2(t)\,. \tag{4.41}$$

where $\hat{v}(t)$, the variance of the degree sequence, is a proxy for $\hat{\gamma}(t)$. These estimators are easy to update on the fly in $O(1)$ operations; one simply needs to keep a running count of $|V_t|, |E_t|, \sum_{i\in V_t} k_i^2(t)$ and t (notice that $\sum_{i\in V_t} k_i(t) = 2t$ by definition). Now, for target values (v^*, b^*) computed on the final graph (i.e., the target baseline), and some control parameters (β, μ), we give the following weight to an edge :

$$w(e) \propto \exp\left\{ -\beta\left[\hat{b}(t+1) - b^*\right]^2 - \mu\left[\hat{v}(t+1) - v^*\right]^2 \right\}. \tag{4.42}$$

In this equation, $\hat{b}(t+1)$ and $\hat{v}(t+1)$ are the estimators evaluated in the hypothetical history where e is selected at time t, We then draw the next edge with probability proportional to $w(e)$.

Much like its unbiased version, the biased snowball sampler naturally works as a proposal distribution for the importance sampling method, because we can compute the probability $Q_{bsb}(X|G;\beta,\mu)$ of generating a particular sample X on the fly, by accumulating the transition probabilities. Specifically, if we denote the edge selected at time t as e_t, and the sum of all weights at time t of history X as $W_X(t) = \sum_{e\in\Omega_X(t)} w(e)$, then

$$Q_{\text{bsb}}(X|G;\beta,\mu) = P(e_0|G;\alpha) \times \prod_{t=1}^{T-1} \frac{w(e_t)}{W_X(t)}\,. \tag{4.43}$$

The unbiased snowball sampler is recovered by setting $\beta = \mu = 0$, since $w(e_t)/W_X(t)$ is then equal to $1/|\Omega_X(t)|$.

We show in Figs. 4.13–4.14 the evolution of the estimators $(\hat{b}, \hat{v})$ in both *unbiased* and *biased* ("controlled") samples generated using the snowball sampling algorithm. We superimpose the trajectories associated with the ground-truth for reference. We use a bias of $\alpha = 5$ for all controlled trajectories (see above discussion), and of $\alpha = 0$ in all uncontrolled trajectories. We choose different values of (β, μ) based on $(\hat{b}(T), \hat{\gamma}(T))$ [3]. If $\hat{b}(T) < 1$ (i.e., if the graph is not a tree), then our experiments suggest that controlling the evolution of the density through β is enough to get good samples with respect to *both* the variance $\hat{v}(t)$ and the ratio $\hat{b}(t)$. Therefore we set $\mu = 0$ and a value of β based on T. For the small graphs of $T = 50$ edges of Fig. 4.13, $\beta = 50$ appears to be enough . For the large graphs of $T = 1000$ edges of Fig. 4.14, we use $\beta = 2000$, since much more drastic corrections are necessary to obtain a plausible $\hat{b}(t)$ curve (see the extreme trajectories of $\hat{b}(t)$ in uncontrolled snowball samples). On trees, β has no effect and we turn to μ. We favor convergence towards the steady state with $\mu = 1$, except

[3] Here we evaluate γ directly—only once on the full graph—to choose values for bias parameters. This is a relatively easy minimization problem, see Materials and Methods.

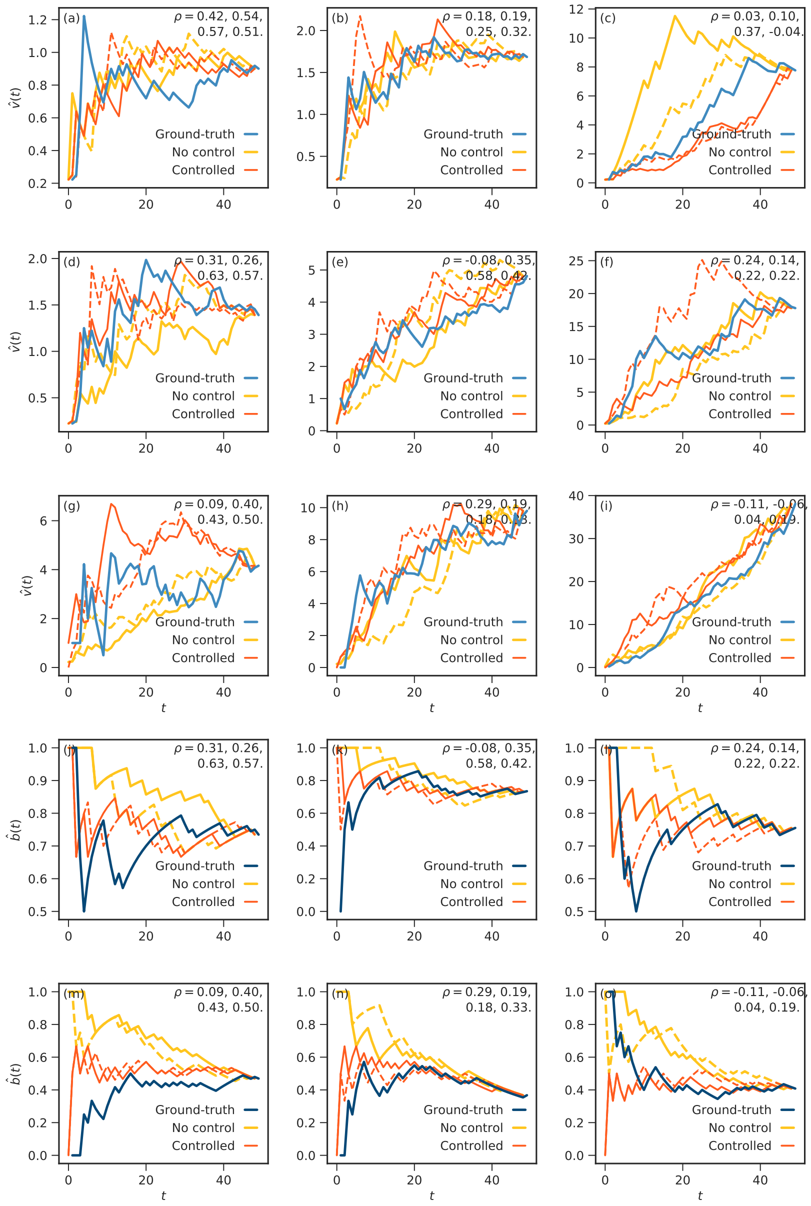


FIGURE 4.13 – Evolution of the ground-truth (blue lines), of two uncontrolled samples (yellow lines), and of two controlled samples (orange lines), on artificial graphs ot $T = 50$ edges generated by the model. We use the kernel exponent (left column) $\gamma = -1$, (central column) $\gamma = 0$, and (right column) $\gamma = 1$, and the node creation probability equals (top row) $b = 1$, (second and fourth row) $b = 0.75$, and (third and last row) $b = 0.50$. The first 9 panels show the evolution of the estimated variance $\hat{v}(t)$ and the last 6 panels show the evolution of the estimated node creation probability $\hat{b}(t)$. The panels are matched, e.g., the dotted orange line in panels (d) and (j) illustrate the evolution of the estimators in a single sample, on the same graph. There are only six panels associated with $\hat{b}(t)$ since its evolution is trivial on trees (i.e, when $b = 1$). The number in the top-right corner of each panel indicates the correlation of the ground-truth and the uncontrolled samples (two topmost numbers), and of the ground-truth and the controlled samples (the two numbers below).

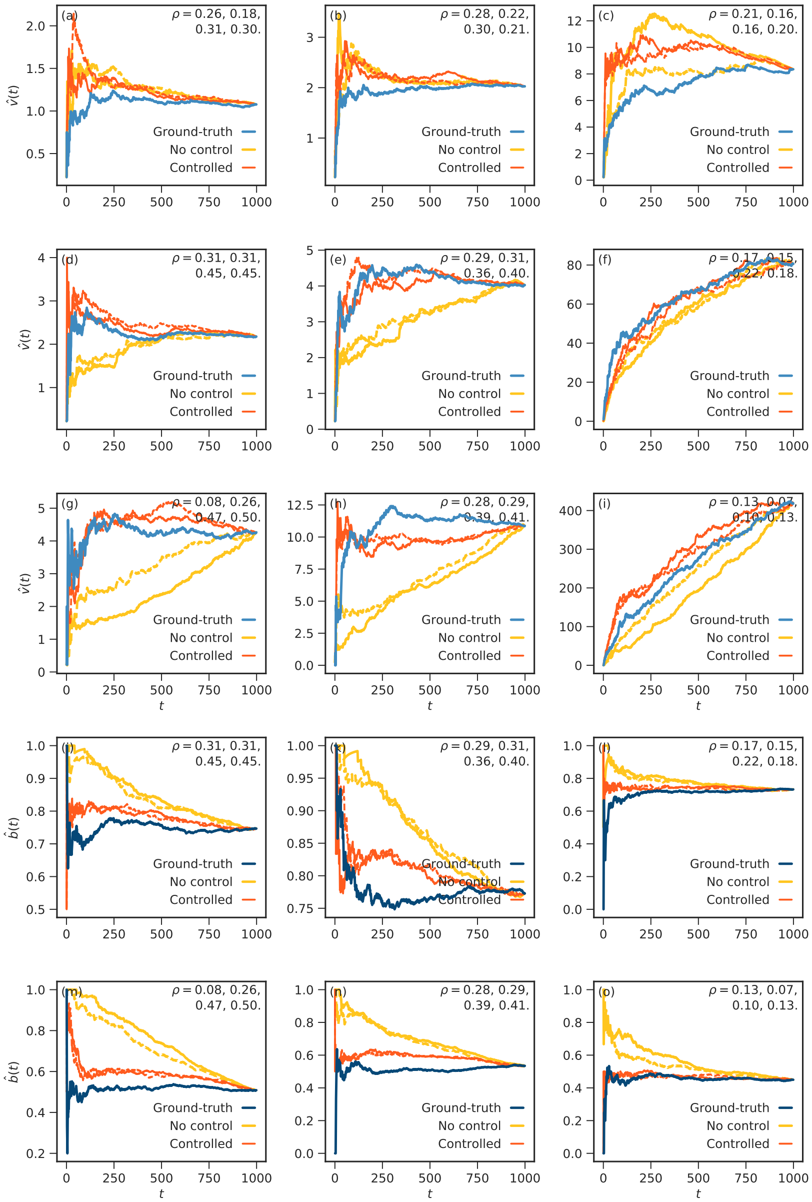


FIGURE 4.14 – Evolution of the estimators in snowball samples of larger graphs. See caption of Fig. 4.13 for a description of how to read this figure. The only crucial differences are that the graphs now comprise $T = 1000$ edges, and we use $(\beta = 2000, \mu = 0)$ whenever $b < 1$. Control plays a much more important role in large graphs.

in the scale-free case ($\hat{\gamma}(T) \approx 1$), where the variance never settles. In this situation we opt not to control the sampler at all, by fear of reaching the steady state of $\hat{v}(t) = v^*$ too rapidly. This illustrates a limitation of our choice of proxy for γ : Unlike γ, the variance goes through a transitive state before the degree distribution of the model settles on its final steady state. With access to $\hat{\gamma}$, we could have perhaps intervened even in the scale free case.

Figures 4.13–4.14 confirm that the biased sampler is significantly more accurate than the unbiased version. One could conclude that, as a result, the biased sampler should be systematically favored. But it turns out that in practice, the biased sampler is not useful at all, because weighting the boundary adds a significant computational overhead. In the unbiased version of the snowball sampler, one can draw an edge from the boundary in $O(1)$ time (by storing the boundary in a container with random access). The limiting step is instead the examination of the neighborhood of the edge selected at time t, of order $O(k_{\max})$ where $k_{\max}$ is the largest degree in the graph. But in the *biased* version of the algorithm, the limiting step becomes the drawing step, of complexity $O\big(|\Omega_X(t)|\big) \gg O(k_{\max})$. This is due to the fact our weighting scheme is both *dynamical* and *global* : All the weights must be recomputed at each step because they are functions of time-dependent quantities (the estimators) that are functions of the whole graph. Therefore, even though the samples of the biased sampler are better, they also come with a higher, impractical, price tag. In our experiments, we have found it more advantageous to generate many poorly correlated samples, than comparatively fewer good samples, in the same amount of time.

The biased sampler is however not entirely useless, because it suggests interesting directions for further research. Indeed, our numerical experiments highlight the fact that a typical snowball sample is not similar to the ground-truth at all. In the early stage of the enumeration, the snowball sampler branches out on the background graph [219], generating an almost tree-like sample. Only in the later stages does it consolidates edges and lowers its node to edge ratio $\hat{b}(t)$ An improved sampler should account for this—like the biased snowball sampler—, but only efficiently so. Obtaining a good *local* weighting scheme would go a long way towards obtaining better efficient samplers, since it could be implemented as a $O\big(\log|\Omega_X(t)|\big)$ drawing step, using weighted binary trees to store the edges of the boundary.

Sampler convergence

The quality of the histories inferred with the MMSE estimators is obviously tied to the number of samples used to compute the averages. This relation is investigated in Fig. 4.15, where we analyze the interplay between the sample size and the achieved correlation. Our experiments show that the importance sampling method converges in roughly $n = 10^6$ samples (for graphs of $T = 50$ edges), in the sense that adding more samples does not appear to increase the average correlation significantly, or to further concentrate the distribution of the correlation on its mean. In all cases the sampling method outperforms the onion decomposition

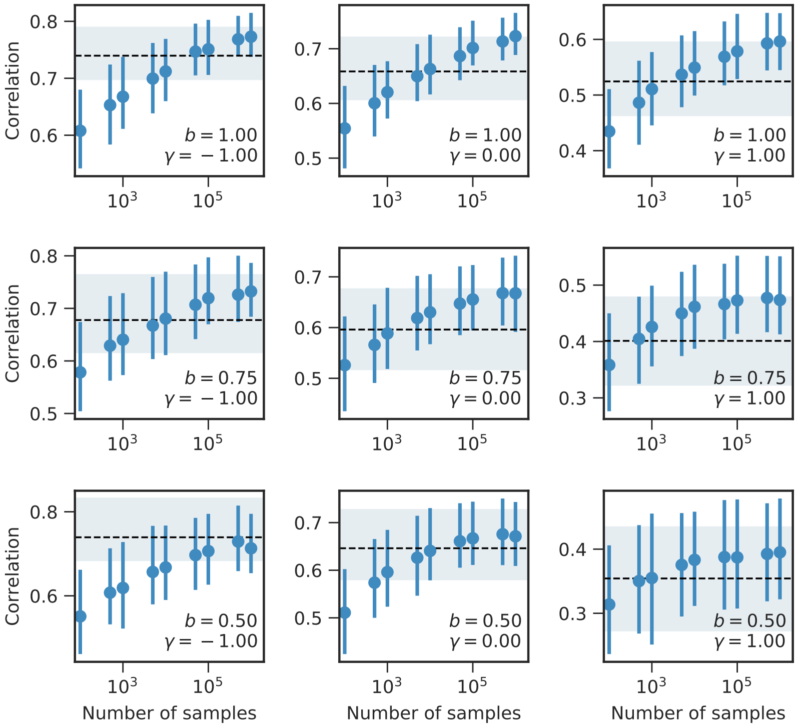


FIGURE 4.15 – Average correlation achieved by the importance sampling algorithm with a snowball proposal distribution, on artificial networks of $T = 50$ edges generated with $b \in \{1, 0.75, 0.5\}$ (from top to bottom), and $\gamma \in \{-1, 0, +1\}$ (from left to right). Data points are obtained from 250 independent realization of the model and of the sampling process (except for $n = 10^6$, where only 100 realizations are used). 50% of the outcomes lie within the error bars (from the 25th to the 75th percentile). The performance of the onion decomposition is shown for reference (average denoted by a dotted black line and the quartiles by the shaded rectangle).

(OD) on average given enough samples. While this sample size is certainly large, $n = 10^6$ samples still represent an astronomically small fraction of the $|\Psi(G)|$ potential histories[4].

These "averaged results", however, should not be taken as a proof of the dominance of the sampling method over the OD on *an instance-by-instance* basis, even given infinitely many samples (see Fig 4.16). Indeed, it is unclear whether the sampling method would systematically outperform the OD given a perfect knowledge of the posterior distribution $P(X|G,\gamma,b)$. The optimality proof of Sec. 4.8.3 only guarantees that the correlation is maximized on ave-

[4] A loose upper bound on $|\Psi(g)|$ is $T!$, not accounting for the existence of inconsistent histories. Closer estimates can be obtained with a small modification to the sequential importance sampling method (see Ref. [24] for a related application of sampling to counting). The method relies on the identity $|\Psi(G)| = \sum_{X\in\Psi(G)} Q(X|G)/Q(X|G) = \langle 1/Q(X)\rangle_Q$. The idea is to approximate the average as $\frac{1}{n}\sum_{i=1}^{n}[Q(x_i|G)]^{-1}$ with x_i drawn from the distribution $Q(X)$, here chosen as the snowball proposal distribution for convenience. This yield a precise estimate of the size of $\Psi(G)$. In the classical preferential attachment case where $b = 1$ and $\gamma = 1$, for example, we find $\langle|\Psi(G)|\rangle = O(10^{54})$ when $T = 50$, a result that is ten orders of magnitude smaller than the loose bound $T! = O(10^{64})$.

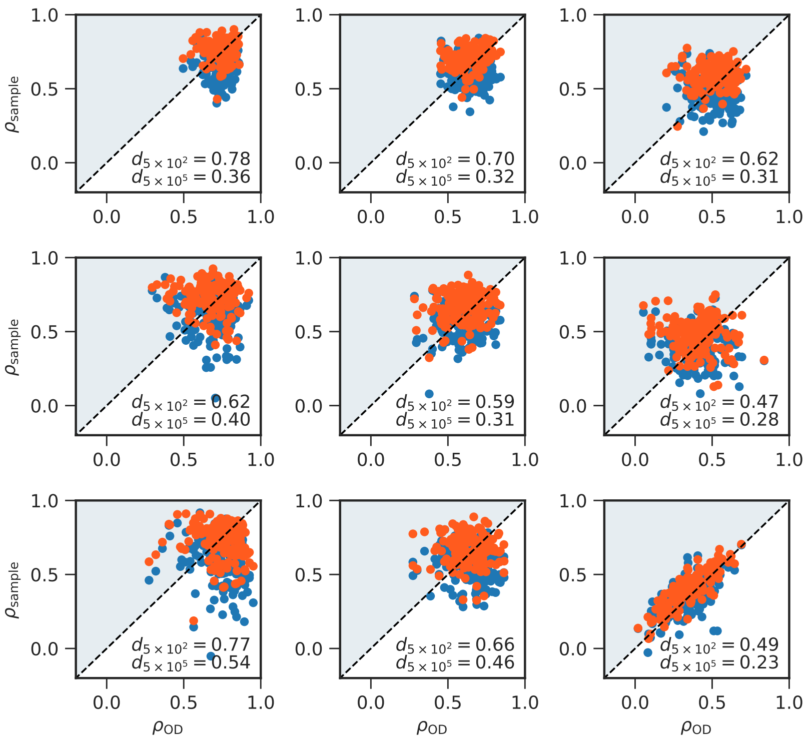


FIGURE 4.16 – Instance-versus-instance comparison of the results appearing in Fig. 4.15. Each dot corresponds to a specific network realization. We show the correlation ρ_{OD} achieved by the onion decomposition versus the correlation ρ_{sample} achieved by the importance sampling method, with 5×10^2 samples (blue) and 5×10^5 samples (orange). The fraction d_n of instances for which $\rho_{\text{OD}} < \rho_{\text{sample}}$ appears in each figure. Points in the shaded region denote instances where the sampling outperformed the onion decomposition.

rage, over the set of all histories consistent with some graph G. In other words, the proof relies on an average notion of optimality, rather than one couched, e.g., in the languages of worst-case guarantees. As a consequence, there could be instances of the model for which the OD works better than the perfect MMSE estimators (or their numerical approximations). It is in fact trivial to construct such instances artificially, by conjuring an arbitrary "ground-truth" perfectly aligned with OD, for some arbitrary graph. The relevant question, then, is whether these instances are *common* or so *unlikely* that they will never be observed in practice. The latter scenario would lead to a practical dominance of the MMSE method. The results of Fig. 4.16 suggest that reality lies somewhere in between : There are indeed some instances where it is hard—perhaps impossible—to outperform the simple OD method, but in the majority of the cases, well approximated MMSE estimators triumph.

This could be better settled by running these experiments with even more samples and independent graphs. The sample size used here is relatively small, due to computational constraints.

4.8.6 Details of the parameter estimation procedure

The statistical inference approach introduced in the main text is developed with the explicit assumption that the parameters of the generative model are either known, or correctly estimated prior to any archeology steps. We here characterize this estimation step, and show that the parameters are indeed recoverable in large instances, by way of simple calculations that do not rely on the full posterior distribution of the model.

Node creation probability

The quantity b determines whether new edges involve two existing nodes (prob. $1-b$) or an existing node and a new node (prob. b). We therefore term it the "node creation probability."

Our estimator of b relies on the observation that a graph G can be seen as a signature of $|E(G)|-1$ i.i.d. Bernoulli trials of parameter b. Indeed, each edge beyond the first embodies a test of whether a new node should be added, and each node beyond the two initial nodes signals a success of the trial. Therefore the classical theory of estimation applies to the node creation probability. We opt for the simple maximum a posteriori (MAP) estimator

$$\hat{b}(G)=\frac{|V(G)|-2}{|E(G)|-1}\,, \tag{4.44}$$

valid in the large graph limit or when the prior on $\hat{b}(G)$ is uniform. Note that upon generalizing the model to different seed graphs G_0 [144], the asymptotic estimator would instead read :

$$\hat{b}(G)=\frac{|V(G)|-|V(G_0)|}{|E(G)|-|E(G_0)|}\,. \tag{4.45}$$

Kernel exponent

The exponent γ of the attachment kernel controls the degree heterogeneity. Our goal is to learn this exponent network structure alone (without using temporal data, such that approaches like that of Ref. [113] do not apply).

Before we introduce our estimation procedure, let us discuss its foundation, the (reasonable) assumption that the degree distribution of the graphs generated by the model is a sufficient statistic of the exponent of the kernel. As far as we know, the sufficiency of the degree distribution is not discussed in the literature, but at least three arguments support it (besides the empirical characterization below). First, a detailed characterization of the behavior of the degree distribution with respect to γ hints that the expected degree distribution is uniquely determined by γ for all $\gamma<2$, with many qualitative transitions at special values of γ, see main text and Ref. [128]. Second, the exponent directly controls the degree heterogeneity ; we can reasonably expect to observe its signature in the degree distribution. Third and finally, while an estimator based on the degree distribution will obviously fail to differentiate two

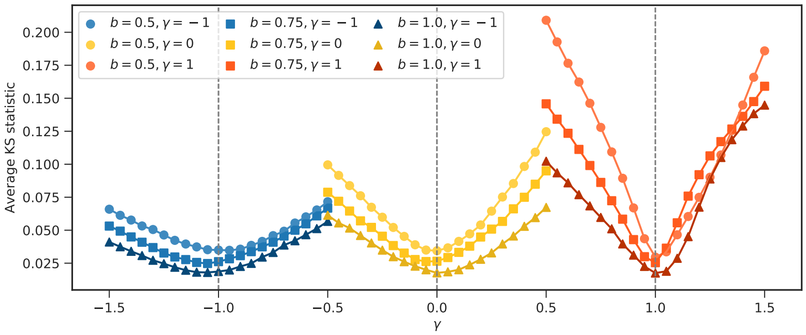


FIGURE 4.17 – Demonstration of the convexity of the average KS–statistic. Each curve is associated with a single random graph, of $T = 1\,000$ edges, with generating parameters (γ^*, b^*) indicated by the color and shape of the markers. For every instance G, we compute the average KS–statistic of the empirical degree distribution $P(G)$ and of $n = 1\,000$ random degree distributions, generated with parameters $(\gamma, \hat{b}(G))$ in the range $\gamma \in \left[\gamma^* - \frac{1}{2}, \gamma^* + \frac{1}{2}\right]$. The figure confirms that the average KS–statistic is convex and centered near the generating parameters (indicated by vertical lines).

parameterizations where $\gamma > 2$, one can argue that these networks have no distinguishing features whatsoever to begin with (they are, more or less, star-like networks) ; *all* estimations are expected to perform poorly.

Given some value of γ and the MAP estimator $\hat{b}$, we quantify the distance between the model and the data using the average Kolmogorov-Smirnov (KS) statistic of the empirical degree distribution $P(G)$ and of the degree distributions of random instance of the model $\{Q^{(i)}(\gamma)\}_{i=1,\ldots,n}$. We obtain our estimator $\hat{\gamma}$ by minimizing this function with respect to γ—a now standard approach [42]. The average KS divergence is convex with respect to γ, such that $\hat{\gamma}$ can be found efficiently via bisection or Brent's method [204] (see Fig. 4.17). The KS–statistic of a pair of noisy distributions (P, Q) is given by the supremum of the difference of their cumulative distribution function (CDF), i.e.,

$$D(P, Q) = \sup_k |f_P(k) - f_Q(k)| \, , \tag{4.46}$$

where $f_P(k)$ is the CDF of P at point k. In all our characterization tests shown in this document, we use $n = 1\,000$ samples to approximate the true average KS–statistic. When fitting real networks, we use $n = 100\,000$ samples, since we need not sweep the whole parameter space.

Because the minimization works correctly only when the KS–statistic is smooth with respect to γ, we need $n \gg 1$ samples to properly approximate it. Unfortunately, simulations become prohibitively expensive in the limit of large networks and large sample sizes. Hence, we opt for a more efficient solution : We first integrate the mean-field equations of the model (see (4.17)), and then draw n finite samples from the distributions. This method is equiva-

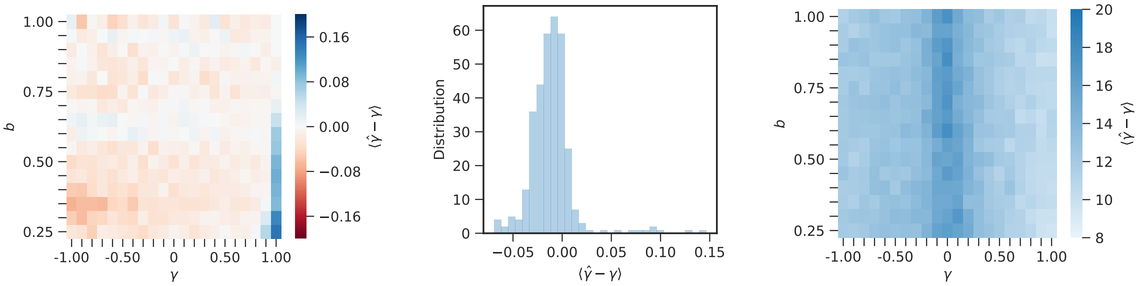


FIGURE 4.18 – (left) Difference between the true value and estimated value of γ, for networks of $T = 1\,000$ edges generated with parameters $\gamma \in [-1, 1]$ and $b \in [0.25, 1.00]$. The difference $(\gamma - \hat{\gamma})$ is averaged over 20 different network realizations at each point of the parameter space. (center) Distribution of the values appearing in the left panel. The error $(\gamma - \hat{\gamma})$ is close to zero on average, with a small systematic negative bias. However, we note that the estimator is strongly biased at $\gamma = 1$ with $b \ll 1$. This effect can be traced back to the poor fit between the model and its mean-field description, see Sec. 4.8.1. (right) Average number of evaluations needed to attain convergence in Brent's method. Exponents closer to 0 are harder to find. We use $n = 1\,000$ samples each time we evaluate the average KS–statistic.

lent to—but much faster than—n repeated simulations. The consistency of the estimators is investigated in Fig. 4.18

Goodness of fit

The estimation framework provides a natural test of the goodness of fit of γ [42]. The procedure goes as follows. After we have found $\hat{\gamma}$ via the minimization of the average KS–statistic, we keep the corresponding minimum KS–statistic $D^*(G)$ in memory. We then generate n_{bs} random degree distributions $Q^{(i)}$ from the model of parameters $(\hat{\gamma}, \hat{b})$ and compute the distribution of their KS–statistic, by comparing each of them against n additional random degree distributions $\{S^{(j)}\}_{j=1,\ldots,n}$. This provides a null distribution for D, which tells us whether $D^*(G)$ is an extreme value of the average KS–statistic or not. See Fig. 4.19 and its caption for an example.

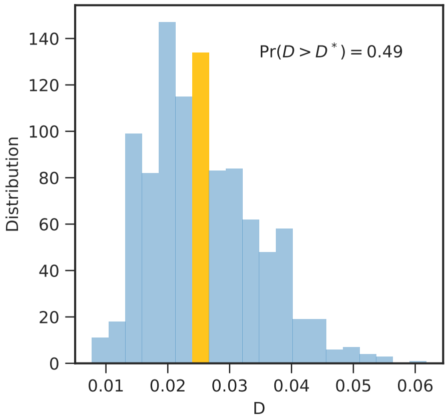


FIGURE 4.19 – Verification of the validity of the test of the goodness of fit (for artificial networks). To ensure that our test of the goodness of fit works properly, we check whether it lends evidence to the generative model when it is applied to a graph generated by the model itself. We use an artificial graph $G = (V, E)$ generated with the parameters $(T = 1\,000, \gamma = 0.7, b = 0.9)$ as our test dataset. We begin by inferring the parameters back, mimicking a typical situation where the source of G is unknown. For the particular graph used to produce the figure, the estimated parameters are $\hat{\gamma} = 0.643$, $\hat{b} = 0.893$, and they are associated with an average KS–statistic of $D^* = 0.025$. We then turn to the test of the goodness of fit [42]. We first draw $n_{bs} = 1\,000$ random degree distributions $\{Q^{(i)}\}_{i=1,\ldots n_{bs}}$ from the model (of parameters $\hat{\gamma}$,$\hat{b}$), and generate $n = 1\,000$ additional degree distributions $\{S^{(j)}\}_{j=1,\ldots,n}$ for each $i = 1, \ldots, n_{bs}$. Then, for each i, we approximate the expected KS–statistic by averaging $D(Q^{(i)}|S^{(j)})\rangle$ over all $\{S^{(j)}\}_{j=1,\ldots,n}$. This last step yields the distribution shown in the figure (with the bin where D^* falls highlighted). Here, the test of the goodness of fit shows that the KS–statistic obtained by minimization is at least as extreme as the one associated with roughly 50% of the samples, providing strong evidence for the model with parameters $(\hat{\gamma}, \hat{b})$, as expected.

4.8.7 email-Eu-core dataset : detailed analysis

The `email-Eu-core` dataset [184] is part of the Stanford Large Network Dataset Collection. It comprises 332,334 emails (directed edges) sent between 986 individuals over a span of 803 days within a large European research institution. Each email is represented by a timestamped, directed edge.

Preprocessing : Method

We apply two preprocessing steps to extract a growing network amenable to generative modeling with our random attachment model.

First, we obtain a simple growing graph by *thresholding* the temporal network : Starting from $t = 0$, we run the dataset forward in time, and count the number of edges appearing between each pair of nodes. A pair of nodes becomes connected by an undirected edge as soon as one of the nodes has sent at least n_c^{up}, and has received $n_c^{\text{low}} \leq n_c^{\text{up}}$ emails in return. For suffi-

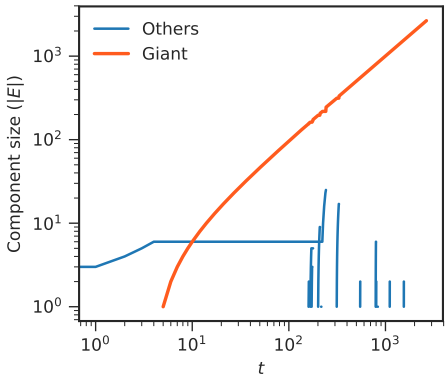


FIGURE 4.20 – Evolution of the largest component for the pre-processed email dataset. Example of how the size of the largest component (solid orange line) evolves in time, here computed on the full email dataset with the thresholds $n_c^{\text{low}} = 5$ and $n_c^{\text{up}} = 10$. Blues lines show the evolution of the size of the smaller components. These components disappear upon joining the giant component. In this example, only 2.9% of all edges are added to the giant components during a merger event.

ciently large values of n_c^{low} and n_c^{up}, this thresholding step removes noise from the dataset, e.g., one-way institutional communications or transient contact. It is then safer to interpret the resulting network as a growing *social* network, where two individuals are connected if they have exchanged enough emails.

Second, we obtain a single growing component by focusing on the history of the giant component alone. This step is needed because according to our version of PA, $P(X|G,\theta) = 0$ for any history that contains disconnected components at some point in time. To extract the history of the giant component from the dataset, we run its evolution backward in time, and make a few simplifying assumptions to deal with the multitudes of components that merge in growing graphs : (i) each time the giant components splits in two (in the reversed time direction), we assume that the largest of the two resulting components is the giant component. (ii) when splitting events occur—again in reverse time—, we assume that the edges of the smaller component leave the giant component sequentially, starting at the time of the split, in an order that parallels their ordering *within* the smallest component. Running this reversed history forward again yields a growing graph consistent with the PA model.

In practice, we find that these events usually involve components of vastly asymmetric sizes, and that they mainly occur in the first steps of the growing process (see Fig. 4.20). As a result, the simplifying assumptions do not have much of a global impact.

Preprocessing : Results

The choice of thresholds $(n_c^{\text{low}}, n_c^{\text{up}})$ obviously influences our ability to infer the time of arrival of edges in the resulting temporal network. For low thresholds, this network is large,

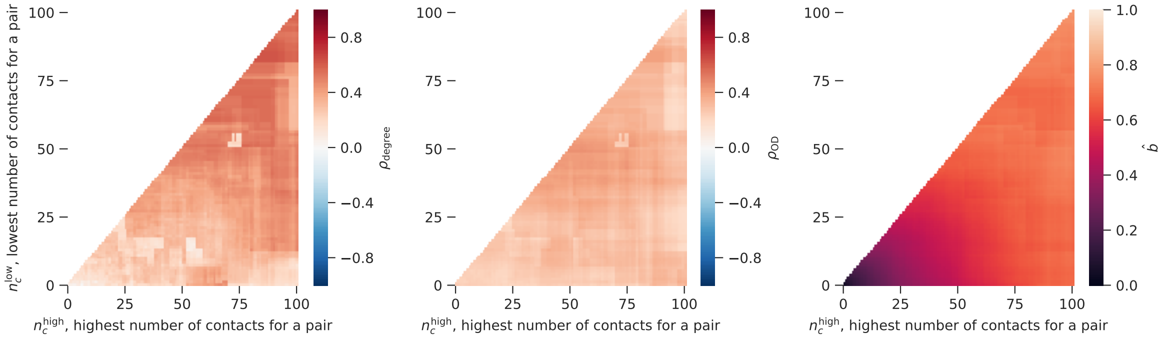


FIGURE 4.21 – Correlation attained by (left) the degree ordering and (center) the onion decomposition, as well as (right) node to edge ratio $\hat{b}$, for various choices of asymmetric thresholds on the temporal email dataset. An undirected edge is added between two individuals as soon as either of them has sent at least n_c^{up} emails to the other and has received $n_c^{\text{low}} \leq n_c^{\text{up}}$ emails in return (vertical and horizontal axis of the figures, respectively). The unthresholded dataset comprises 332 334 emails between 986 individuals sent over 803 days [184].

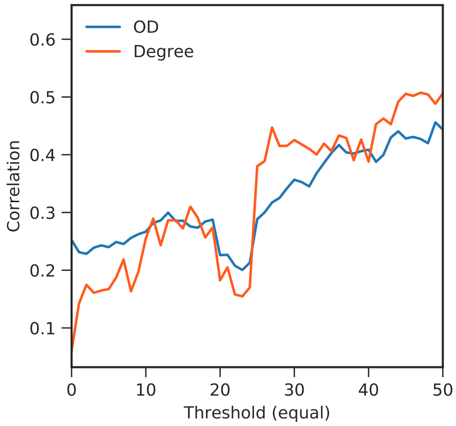


FIGURE 4.22 – Correlation as a function of an equal threshold $n_c^{\text{low}} = n_c^{\text{up}}$ (diagonal of Fig. 4.21).

dense and noisy; inference is more or less impossible. For high thresholds, this network is extremely sparse and uninteresting; inference is possible, but somewhat artificial. Our goal in thresholding is to strike a balance between these two extremes.

As a rule of thumb, both thresholds should be large enough to filter out noise. Otherwise we risk categorizing one-way interactions (e.g., advertisements for a seminar) as likely social contacts. The thresholds should however not be so high as to exclude casual relationships (e.g., a few emails between the members of a committee). With this in mind, thresholds where both n_c are of order $O(10^1)$ appear suitable, considering that the dataset only spans 803 days.

We put these observations to the test in Figs. 4.21–4.22, where we run simple inference me-

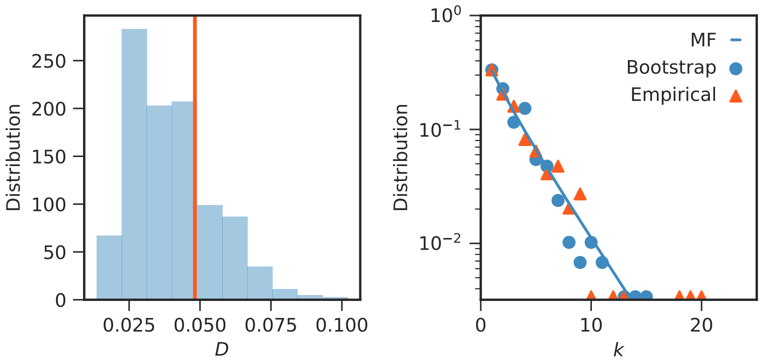


FIGURE 4.23 – Goodness of fit of the degree distribution (for the email network). (left) Null distribution for the average KS–statistic. The value D^* obtained for the best fit to the real dataset is indicated with an orange vertical line. This distribution is estimated with $n_{bs} = 1000$ random degree distribution, each compared against $n = 100$ additional distributions to compute the average KS-statistic. (right) Degree distribution the network (orange triangle) compared with the mean-field distribution of the best fit (solid line) and one bootstrap samples of this distribution (blue circles).

thods on the giant component of many thresholded networks. The onion decomposition and the degree-based inference algorithm both identify histories that are positively correlated with ground-truths, regardless of the choice of thresholds. As predicted, increasing the threshold removes noise and leads to more robust predictions. The two methods behave similarly on the range $n_c = 0, \ldots, 50$. Larger thresholds lead to networks that are extremely sparse and small, where inference is impossible.

As argued in the main text, the model is only a good fit for a network if $\rho_{\mathrm{OD}} > \rho_{\mathrm{degree}}$ (or at least two close results), because that is the observed ordering of correlations on artificial networks generated by the model. With this in mind Fig. 4.22 suggests that a symmetric threshold of $n_c \approx 40$ is suitable (there are no principled motivations for preferring an asymmetric threshold). This choice is also motivated by the facts that the generative model is only a sensible approximation for simple graphs if b is not too small (we get $b = 0.63$ at $n_c = 40$); otherwise we would expect many self-loops and parallel edges.

Parameter estimation and inference results

We submit the filtered network to the parameter estimation procedure presented in Sec. 4.8.6 of this Supplementary Material. We find that it is best modeled by $(\hat{\gamma}, \hat{b}) = (0.0714, 0.6290)$, associated with an average KS-statistics of $D^* = 0.0399$. Using $n_{bs} = 1000$ random degree distributions, each compared against $n = 100$ additional distributions, we find that the $D*$ is significant, with $P[D > D^*] = 0.24$ [42]. (see Fig. 4.23). Running the inference methods on the full network, we obtain the correlations $\rho_{\mathrm{Degree}} = 0.388$, $\rho_{\mathrm{OD}} = 0.409$ and $\rho_{\mathrm{MMSE}} = 0.615$ (with only 10^4 samples).

Épilogue

Chapitre 5

Au-delà des interactions binaires

Article original :

"Construction of and efficient sampling from the simplicial configuration model"

Jean-Gabriel Young [1] , Giovanni Petri [2] , Francesco Vaccarino [2, 3], Alice Patania [2, 3]

[1] Département de Physique, de Génie Physique, et d'Optique
Université Laval, Québec (QC), G1V 0A6, Canada

[2] ISI Foundation, Torino, 10126, Italy

[3] Dipartimento di Scienze Matematiche, Politecnico di Torino, Torino, 10129, Italy

Référence : Phys. Rev. E **96**, 032312 (2017) [250]

‡ Ces sections sont reproduites directement de l'article orignal. Le contenu n'en a été modifié que pour se conformer au format exigé par la Faculté des études supérieures et postdoctorales de l'Université Laval.

5.1 Avant-propos

Dans les chapitres précédents, nous avons accepté, implicitement, une hypothèse forte : à savoir que la structure des systèmes complexes peut être réduite à une collection d'interactions binaires.

Or, il est depuis longtemps entendu que la structure de certains systèmes n'est pas forcément bien représentée par un réseau classique. Considérons par exemple un système où les interactions impliquent plusieurs corps (e.g. interactions moléculaires en biologie [81]). On peut en principe en faire une représentation réseau, en ayant recourt à un «*line graph* » (LG) [94]. Il suffit d'introduire deux ensembles de noeuds : L'ensemble des éléments du système V (les noeuds à proprement parler), et un ensemble de noeuds fictifs F représentant les interactions. On connecte ensuite les noeuds de V via F. Cette représentation est certes puissante et universelle, mais elle n'est pas forcément la plus appropriée et *naturelle*, particulièrement si on ajoute des contraintes supplémentaires sur le système. Par exemple, si on demande que les interactions respectent une contrainte d'inclusion de la forme $X = \{a,b,c\} \in F \implies \mathcal{P}(X) \in F$ [où $\mathcal{P}(X)$ est l'ensemble puissance de X], alors le LG devient rapidement trop grand pour être manipulé numériquement.

Il s'agit d'un réel problème, car les contraintes d'inclusions apparaissent régulièrement dans les données empiriques [1]. On se doit donc de développer des outils numériques et mathématiques pour les traiter. Dans cet épilogue, on utilise une représentation par *complexe simplicial* pour analyser ce genre de données, car plusieurs développements récents suggèrent qu'il s'agit de la méthode toute indiquée [183, 186].

5.2 Résumé

Les complexes simpliciaux offrent maintenant une alternative populaire aux réseaux lorsque vient le temps de décrire la structure d'un système complexe, principalement parce qu'ils peuvent encoder des interactions multiples explicitement. À terme, cette nouvelle représentation se devra d'être accompagnée d'outils statistiques rigoureux—par exemple des modèles nuls—permettant l'étude de données empiriques complexes. Dans ce chapitre, on propose un modèle nul naturel pour un complexe simplicial générique, soit le modèle simplicial des configurations (*simplicial configuration model*, SCM). Le coeur de notre contribution est un algorithme Monte Carlo par chaîne de Markov permettant d'échantillonner le modèle uniformément et efficacement. On démontre l'utilité du SCM en tant que modèle nul en comparant la topologie de trois systèmes réels à la topologie moyenne des ensembles nuls associés, via les nombres de Betti de leurs groupes d'homologie. Dans deux cas sur trois,

[1] C'est par exemple le cas lorsque les noeuds d'un système vivent dans un espace métrique, et que leurs interactions prennent la forme d'une fonction de seuil [161].

le modèle nul nous permet de rejeter l'hypothèse qu'il n'existe pas de structure au-delà du niveau local.

5.3 Abstract

Simplicial complexes are now a popular alternative to networks when it comes to describing the structure of complex systems, primarily because they encode multinode interactions explicitly. With this new description comes the need for principled null models that allow for easy comparison with empirical data. We propose a natural candidate, the *simplicial configuration model*. The core of our contribution is an efficient and uniform Markov chain Monte Carlo sampler for this model. We demonstrate its usefulness in a short case study by investigating the topology of three real systems and their randomized counterparts (using their Betti numbers). For two out of three systems, the model allows us to reject the hypothesis that there is no organization beyond the local scale.

5.4 Introduction

Network science's approach to complexity rests onto the tacit hypothesis that the structure of complex systems is reducible to the pairwise interaction of their constituents. It is often a valid premise and, as a result, network science has been extremely successful in, e.g., both predicting [185] and controlling [142] the behavior of complex systems, inferring their function from their structure [171, 203], and so on. Networks, however, might not be as ubiquitous as previously thought. It has been shown recently that the structure of a number of complex systems, such as the brain [52, 85], protein interactions [243] and social systems [102, 228], cannot always be reduced to the sum of pairwise interactions. For these systems, it is now known that network representations can give an incomplete picture : When many-body interactions are broken down into multiple pairwise interactions (cliques), high-order information simply disappears [259].

Simplicial complexes generalize graphs by encoding many-body interactions explicitly ; they have hence been proposed as a complementary description of the structure of complex systems [51, 67, 107, 186]. Different from hypergraphs, they are equipped with an implicit notion of containment. If nodes $(v_1, \dots, v_{q+1})$ are involved in a q-dimensional interaction, then it is implicit that all possible lower dimension interactions involving the same nodes also exist [for example $(v_1, \dots v_q)$ and (v_1, v_3)]. While it might appear constraining, this property actually arises in all systems where interactions are maximal, e.g., in scientific collaborations (largest cohesive group of collaborators) or gene activation pathways (largest group of collectively activated genes). Furthermore, it is found in many processed relational datasets, e.g., in *clique complexes*, obtained by mapping the cliques of networks to simplices [199, 223], or in filtered simplicial complexes [107]. Simplicial complexes thus offer a natural and compact description of the structure of complex systems, both when high-order structures are explicitly available, or when they are extracted from low-order information.

This application of simplicial complexes has led to promising discoveries : We now better understand, for instance, how to detect large viral recombination events [38], how brain networks reorganize under drugs [198], and how the atomic structure of amorphous solids is hierarchically organized [104]. It has become crucial to establish the statistical significance of these findings, a task for which random null models will be needed. There is already a rich and growing literature on random simplicial complexes and topology, ranging from simplicial generalization of Erdös-Rényi models, amenable to analytical treatment [46, 115], to equilibrium formulations of simplicial complex ensembles [47, 259], and growth models that reproduce various emergent patterns observed in real systems [22, 242]. However, null models—in the sense of network science—are still wanting [79, 182].

We address this issue by refining a recently proposed generalization [47] of the (simple) configuration model of network science [79, 156, 178], which we dub the simplicial configuration model (SCM). Different from Ref. [47], we think of our model as a null hypothesis for real systems; we therefore develop a numerical and statistical toolbox instead of focusing on closed ensemble averages. This entails a number of interesting results : One, we define the first simplicial configuration model able to describe arbitrary complexes, in line with our goal of obtaining a generic null model (Sec. 5.5). Two, we propose and analyze an efficient and rigorous sampling algorithm for this model (Sec. 5.6). Three, we use the model to investigate real datasets and show—now using sound statistical arguments—that the local structure of these systems does not always explain their mesoscale structure (Sec. 5.7). We conclude by listing a few important open problems.

5.5 Simplicial configuration model

Informally, a labeled simplicial complex K is the high-order generalization of a network. Formally, it is a collection of simplices incident on a node set $V = \{v_1, \ldots, v_n\}$ [95]. A q–dimensional simplex—the generalization of an edge—is a tuple of $q+1$ distinct nodes denoted $(v_1, \ldots, v_{q+1})$; we say that this simplex is incident on $v_1, \ldots, v_{q+1}$. All simplices not included in a larger simplex are called the *facets* of the complex, whereas a contained simplex is called a *face*; e.g., if K comprises of $\sigma = (v_1, v_2)$ and $\tau = (v_1, v_2, v_3)$, then σ is a face of facet τ. It is always assumed that if facet $\sigma = (v_i, \ldots, v_j)$ is in the simplicial complex K, all elements in the power set of σ are also in K. Therefore, faces need not be enumerated : The structure of a simplicial complex is fully specified by the list of its facets.

Departing from other recent contributions [47], we define the degree d_i of a node v_i as the number of facets incident on v_i and the size s_i of a facet σ_i as the number of nodes it contains (its *dimension* plus one). This local information is summarized by the sequences

$\boldsymbol{d} = (d_1, \ldots, d_n)$ and $\boldsymbol{s} = (s_1, \ldots, s_f)$, where n is the number of nodes and f is the number of facets.

With these notions in hand, we define the simplicial configuration model (SCM) as the uniform distribution over all labeled simplicial complexes with degree sequence $\boldsymbol{d}$ and facet size sequence $\boldsymbol{s}$. In other words, if $\Omega(\boldsymbol{d}, \boldsymbol{s})$ is the set of all labeled simplicial complexes with joint sequences $(\boldsymbol{d}, \boldsymbol{s})$, then the SCM places a probability

$$\Pr(K; \boldsymbol{d}, \boldsymbol{s}) = 1/|\Omega(\boldsymbol{d}, \boldsymbol{s})| \tag{5.1}$$

on K if it has sequences $(\boldsymbol{d}, \boldsymbol{s})$, and 0 otherwise. The model of Ref. [47] is recovered by setting the size of all facets to a constant s.

This particular choice of definition for the SCM is natural for three reasons. First, the SCM directly generalizes the simple CM of network science [79] ; when $s_i = \{1, 2\}$ for all facets, one recovers a graph ensemble with degree sequence $\boldsymbol{d}$. Second, the SCM does not include any correlation—the structure is maximally random beyond the local level. This is reminiscent of the equivalent network model. Third, the SCM can describe the local structure of any simplicial complex, since it allows for arbitrary degree and size sequences. This property is *not* common to all random models of simplicial complexes, for good reasons ; many models are constructed with a focus on the calculation of closed-form expression for a few properties (e.g., the asymptotic entropy) [47, 115, 259]. This commends simplifying assumptions, e.g., a regular facet size sequences [47]. Our definition of the SCM forgoes these simplifications to accommodate arbitrary local structures, at the expense of analytical tractability.

5.6 Efficient sampling algorithm

5.6.1 Constraints on the support

For the SCM to be of any use, one needs to be able to sample from it. This is far from a trivial problem, because there are numerous constraints on the support of the model. It will be easier to see these constraints by first switching to the equivalent *graphical* representation of simplicial complexes.

In this representation, facets are replaced by nodes (we denote by F this new node set, and by $V \cup F$ the complete node set), and an edge connects facet $\sigma_i \in F$ to node $v_j \in V$ in B if and only if σ_i is incident to v_j in K (see Fig. 5.1). Because B encodes the structure of K without ambiguity, we can think of the model in terms of either representation.

As such, one could be tempted to assume that sampling from the SCM of parameters $(\boldsymbol{d}, \boldsymbol{s})$ is equivalent to uniformly sampling from all bipartite graphs with these degree sequences—a solved problem [154]. But this would be wrong : The mapping is not bijective. This is, in fact, where the constraints on the support of the SCM become apparent [47]. Let us introduce

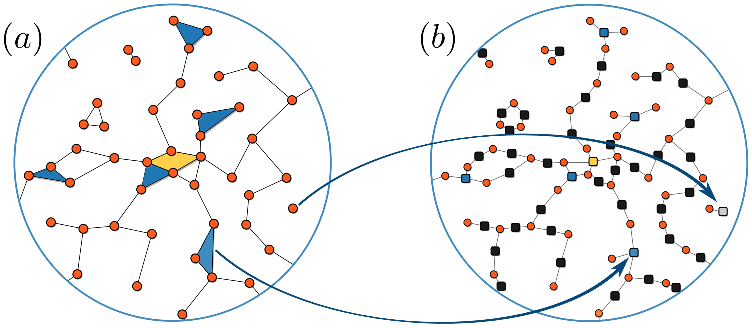


FIGURE 5.1 – (a) Simplicial complex K and (b) its graphical representation B. In the bipartite graph, small square nodes represent facets and large orange nodes represent the nodes of K. An edge connects a facet σ_i and a node v_j in B if the facet σ_i is incident on node v_j in K. Notice how some cliques are not filed (i.e., k fully connected nodes do not necessarily form a size k), and how isolated nodes are attached to facets of size 1.

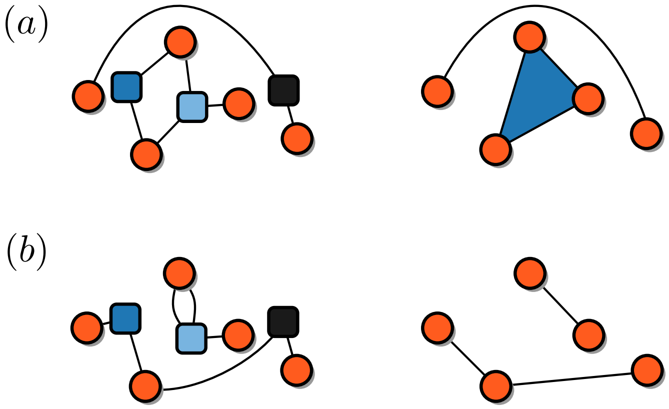


FIGURE 5.2 – Example of non-sequence-preserving bipartite graphs. The two bipartite graphs (left column) encode the joint degree sequences $(\boldsymbol{d}, \boldsymbol{s}) = ([2,2,1,1,1], [3,2,2])$, but the associated simplicial complexes (right column) have different size and degree sequences, respectively $(\boldsymbol{d}', \boldsymbol{s}') = ([1,1,1,1,1], [3,2])$ and $(\boldsymbol{d}', \boldsymbol{s}') = ([2,1,1,1,1], [2,2,2])$. The disparities are due to the presence of (a) a fully included neighborhood, and (b) pairs of nodes connected by more than one edge.

the notion of sequence preserving bipartite graph to formalize these constraints. We say that a bipartite graph B with joint degree sequences $(\boldsymbol{d}, \boldsymbol{s})$ is sequence preserving if, upon interpretation of B as a simplicial complex, one obtains a simplicial complex with facet size sequence $\boldsymbol{s}$ and generalized degree sequence $\boldsymbol{d}$.

Not all bipartite graphs are sequence preserving, and there are two reasons for this, both related to the fact that we think of the nodes in F as facets. The first reason is the inclusion of at least one facet : If there is a $\sigma_i \in F$ such that the neighborhood of $\sigma_i \subseteq \sigma_j$ for some $j \neq i$, then B is not sequence preserving [see Fig. 5.2 (a)]. When this occurs, σ_i is included in σ_j; the corresponding simplicial complex is thus either ill specified (facets cannot contain other facets, by definition) or does not have the same degree and size sequences as B (if we simply remove σ_i). For similar reasons, if two or more edges connect the same pair of nodes in B, then the graph is not sequence preserving [see Fig. 5.2 (b)].

The sampling space would not be too constrained if these non-sequence-preserving bipartite graphs were rare. Sampling would then be easy. Unfortunately, it is straightforward to show that non-sequence-preserving graphs are far more common than sequence–preserving ones, by adapting the calculations of Ref. [20]. We find that the fraction ϕ of bipartite graphs with degrees $(\boldsymbol{d}, \boldsymbol{s})$ not featuring parallel edges rapidly tends to

$$\phi = e^{-\frac{1}{2}\left(\langle d^2\rangle/\langle d\rangle - 1\right)\left(\langle s^2\rangle/\langle s\rangle - 1\right)}, \tag{5.2}$$

where $\langle x^k \rangle$ is the kth moment of the sequence $\boldsymbol{x}$, and where it is assumed that the elements of $\boldsymbol{d}$ and $\boldsymbol{s}$ do not grow with n (i.e., B is sparse). Thus, based on the presence of multi-edges alone, there is a stringent upper bound on the fraction of bipartite graphs that are actually in the support of the SCM. An even smaller fraction remains after the bipartite graph with included neighborhood are removed.

5.6.2 Markov chain Monte Carlo method

To sample from the SCM, then, one needs to sample uniformly from a very constrained space, i.e., that of all sequence–preserving bipartite graphs with joint degree sequence $(\boldsymbol{d}, \boldsymbol{s})$. Previously proposed approaches such as rejection sampling do not work well [47], because natural proposal distributions (e.g., stub matching) give an appreciable weight to non-sequence-preserving bipartite graphs [see Eq. (5.2)]. Thus, we turn to the Markov chain Monte Carlo (MCMC) sampling strategy [2], which has been used with great success for the CM [79, 154]. The general idea is to construct a random chain of sequence–preserving bipartite graphs $B_0, \ldots, B_T$, to sample from this chain at regular intervals, and to treat the samples as if they had been drawn identically and independently from the ensemble. The algorithm will be correct if the chain is ergodic (time averages equal ensemble averages) and uniform (all

[2] We provide a reference `c++` implementation of the sampler as well as tutorials at https://www.github.com/jg-you/scm

non isomorphic B are represented equally). These properties are determined by the allowed transformations $B_t \to B_{t+1}$ and the resulting transition matrix $\boldsymbol{\pi}$, where π_{ij} is the probability that B_j follows B_i in the chain. If the move set *connects the space* and the chain is *aperiodic*, then the chain will be ergodic. If the transition matrix is *doubly stochastic* (all rows and columns sum to 1), then the chain will be uniform.

We claim that the following set of moves satisfies all three conditions. Consider L, a random variable on the support $\mathcal{L} = \{2,3,\dots,L_{\max}\}$, where $L_{\max}$ is a parameter and the distribution $\mathbb{P}[L=\ell]$ is arbitrary but non zero everywhere on $\mathcal{L}$ (for illustration purposes, we will use $\mathbb{P}[L=\ell;\lambda] \propto e^{-\lambda\ell}$). At each step of the chain, we pick L edges in B (uniformly at random). We cut these edges and randomly match the stubs stemming from facets to the stubs stemming from nodes. If this matching generates a sequence–preserving bipartite graph B', then we accept the move; otherwise we resample B. This set of moves is similar to the double-edge swap commonly used in graph MCMC [79]. The only difference is the variable number of rewired edges, added to help the sampler better navigate the constrained support [154]. Much like its graphical counterpart, the resulting MCMC algorithm is efficient since drawing L edges and checking for resampling can be done in polynomial times.

The chain is aperiodic because the above set of moves yields a doubly stochastic transition matrix for any distribution $\mathbb{P}$: The total number of possible transitions at each configuration is a constant independent from the configuration considered (resampling guarantees this) [79]. The chain is also aperiodic, because there exists orbits of period 1 (resampling steps) and 2 (all moves are reversible) for any nontrivial $(\boldsymbol{d},\boldsymbol{s})$.

This leaves open the question of whether the support of the SCM is connected by the set of moves or not. We argue that it is, for all $L_{\max} \geq L^*_{\max}$, where $L^*_{\max}$ is bounded by

$$L^*_{\max} \leq 2 \max \boldsymbol{s} \; . \tag{5.3}$$

To prove this, one would have to show that given two sequence–preserving bipartite graphs B_1 and B_2, it is always possible to find a B_3 such that $\big|\Delta_+\big(K(B_1),K(B_2)\big)\big| \geq \big|\Delta_+\big(K(B_1),K(B_3)\big)\big|$, where Δ_+ is the set of facets in $K(B_2)$ that are not in $K(B_1)$, and Δ_- is the set of facets in $K(B_1)$ that are not in $K(B_2)$ [$K(B)$ is the simplicial complex associated to the graph B]. Although a general proof remains elusive, we propose the following non-rigorous argument, valid for sparse simplicial complexes (simplicial complexes with bounded $\max \boldsymbol{d}$ and $\max \boldsymbol{s}$ in the limit $n \to \infty$).

To construct B_3, we first select a facet σ in Δ_+ (incident on the set of nodes Σ in B_2). The conservation of sizes and degrees guarantees that there exists a facet $\tau \in \Delta_-$ of the same size. The idea is then to start from B_1, cut all edges attached to τ and one edge from every node in Σ, match the stubs of σ to those of $v \in \Sigma$, and finally match the remaining orphaned stubs. This algorithm ensures that B_3 is closer to B_2 than B_1 was, because it removes facets from $\Delta_\pm$ (and does not add new facets either : Each $v \in \Sigma$ has at least on facet in Δ_- by the

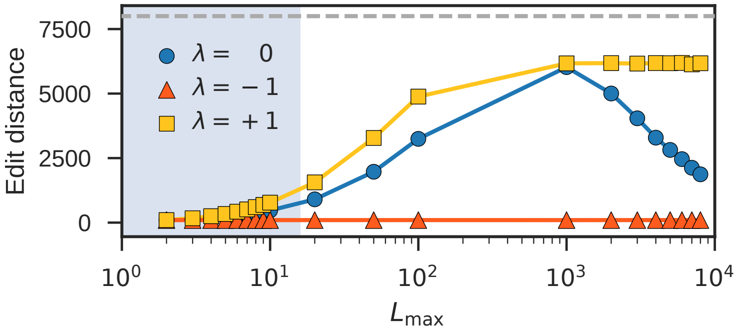


FIGURE 5.3 – Effect of the parametrization of the proposal distribution $\mathbb{P}$ on the mixing time, as quantified by the edit distances of the graphical representation of the samples. We investigate the family of distributions $\mathbb{P}[L = \ell; \lambda] = e^{\lambda \ell}/Z$, and use the regular SCM of $f = 1000$ facets of size $s = 8$, and $n = 2000$ nodes of degrees $d = 4$. Pairs of samples are separated by 100 proposed MCMC moves, and are obtained from a unique initial configuration found via rejection sampling. The shaded region lies below the upper bound on $L^*_{\max}$ of Eq. (5.3). $\lambda = 1$ balances high-rejection probability but efficient moves with safe but inefficient moves, yielding the best overall performance for all $L_{\max}$. In practice, we have found that medium values of $L_{\max}$ are better, because checking for resampling is of complexity $O(L_{\max}\langle d \rangle)$, which translates into slower effective mixing time when $L_{\max} \gg 1$.

conservation of degrees). In general, it is not guaranteed that the last step can be carried out without creating included faces. However, in sparse simplicial complexes, σ is well separated from τ for almost all (σ, τ), since B_1 is locally treelike [178]. In such cases, no included faces are created at the last step, and the above algorithm can be carried through for some (σ, τ), generating B_3. Because this scheme involves at most $L^*_{\max} = 2 \max \boldsymbol{s}$ rewired edges (when $|\tau| = |\sigma| = \max \boldsymbol{s}$), we obtain the bound of Eq. (5.3) for infinite sparse SCM. In practice, $L_{\max} = 2$ seems to always connect the space (we found no counterexamples), and sampling is more efficient when $L_{\max} \gg 2$ (see Fig. 5.3)—the value of $L^*_{\max}$ is thus more of theoretical than practical interest.

5.7 Null model

We put our efficient MCMC algorithm to the test, by verifying the statistical significance of the structural patterns found in three relational datasets that can be represented as simplicial complexes (see caption of Fig. 5.4 for details).

Since an instance of the SCM is provided in each case (the real system), we use it as the initial condition for each independent run of the sampling algorithm. Ergodicity implies that the state of the sampler will be uncorrelated with the initial configuration after a sufficiently long burn-in period—the choice of initial condition is ultimately irrelevant. Extrapolating from the results of Fig. 5.3, we opt for the proposal distribution $\mathbb{P}[L = \ell] = e^{\lambda \ell}/Z$ with $\lambda = 1$

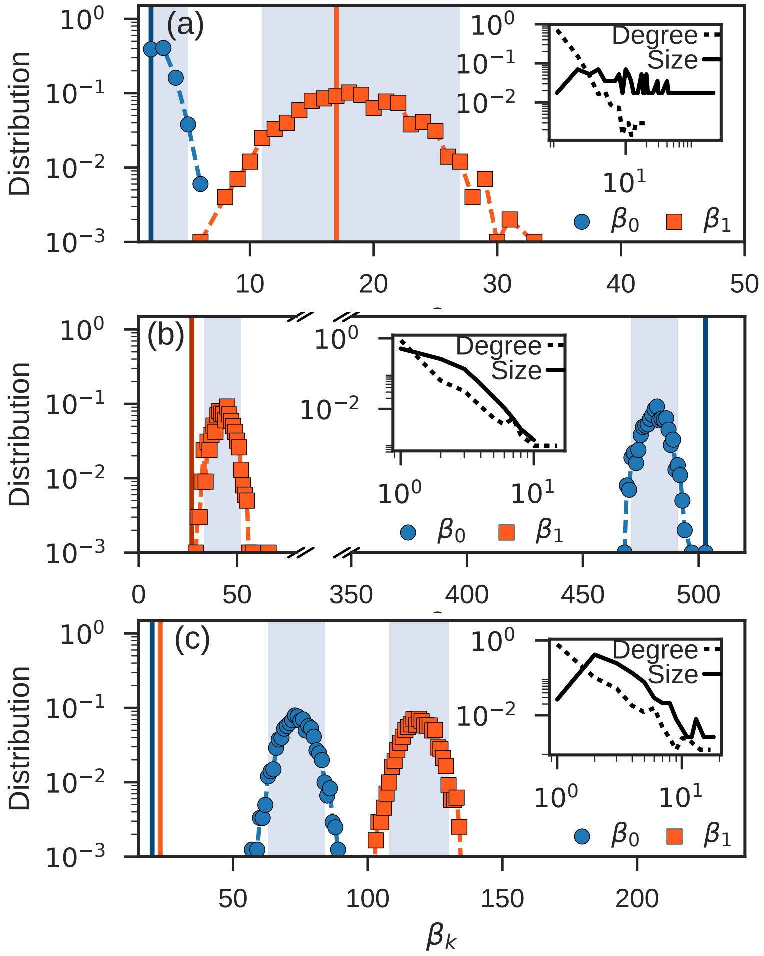


FIGURE 5.4 – Significance of the Betti numbers of real systems. The datasets are bipartite networks, which we convert to simplicial complexes (we prune included faces). They map the relationships between (a) flower-visiting insects (nodes, $n = 679$) and plants (facets $f = 57$) in Kyoto [117], (b) human diseases (nodes $n = 1100$) and genes (facets $f = 752$) linked by known disorder–gene associations [86], and (c) crimes (nodes, $n = 829$) and suspects, victims and witnesses (facets, $f = 378$) in St. Louis [57]. The Betti numbers of these real systems appear as solid vertical lines, and are equal to (a) $\beta_0 = 2$, $\beta_1 = 17$ (b) $\beta_0 = 503$, $\beta_1 = 27$, and (c) $\beta_0 = 20$, $\beta_1 = 23$. We show the distributions of Betti numbers for the equivalent SCM with solid symbols (computed from 1000 instances of the model). The shaded regions contain 95% of the samples. The parameters of the SCM—extracted from real systems—are shown in insets.

and $L_{\max}$ set to 10% of $m = \sum d_i = \sum s_i$. Non rigorous arguments from expander graph theory suggest $t_f = O(m \log m)$ as a good—if overzealous—choice of sampling interval [3].

Significance results only make sense if they rely on a null model that embodies a natural null hypothesis for the problem at hand [79]. For example, the regular CM and its correlated variants usefully show that the network projection of datasets with high-order interactions are abnormally clustered [167]. Therefore, we use the sampler to investigate the distribution of a mesoscopic property only accessible when the datasets are encoded as simplicial complexes : The *shape* of the datasets, as captured by their homology, i.e., the pattern of holes, cavities and higher dimensional voids [95]. The homology can be summarized by a series of Betti numbers $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \ldots)$, where β_k counts the number of structural holes bounded by k-dimensional simplices. For example, β_0 counts the number of connected component, β_1 the number of homological cycles in K, β_2 the number of holes enclosed by facets of sizes 2, etc. Since every instance of the SCM has the same fixed local structure but is otherwise maximally random, we expect significant differences between the Betti number $\boldsymbol{\beta}$ of an organized simplicial complex and the bulk of the distribution of $\boldsymbol{\beta}$ in the corresponding randomized ensembles.

We show in Fig. 5.4 the distribution of β_0 and β_1 for the SCM associated to the real systems. Looking first at β_0, we find that the structure of the pollinator dataset is essentially random [Fig. 5.4(a)]. That is, the overwhelming majority of simplicial complexes with the same sequences have similar β_0. In contrast, the β_0 of the disease genome regulation (hereafter *diseasome*) and crime complexes are highly significant [Fig. 5.4 (b)–(c)] : A random instance of the SCM has fewer (diseasome) or more (crime) components than the real system with high probability. In one case (crime), the difference is a statistical signature of how the dataset was gathered, namely by looking up the ties of suspects, victims and witnesses already in the dataset, recursively [57]. Because this process creates much larger connected components than random sampling, the resulting β_0 is far from the ensemble average—an effect that we expect to find in any dataset constructed using a similar methodology. In the other case (diseasome), the real system has *more* components than one would typically expect from the local information alone. The construction procedure does not explain this disparity [86], meaning that the system must self-organize in a fragmented way, likely for biological or evolutionary reasons.

Turning to β_1 we again find that the structure of the pollinator dataset is typical, and that the same cannot be said of the diseasome and crime datasets. Both simplicial complexes have significantly fewer cycles than expected ; i.e., given a cycle, it is more likely to be filled by a

[3] We represent the support of the SCM as a graph $\mathcal{G}(L_{\max}) = (V, E)$. If $\mathcal{G}$ is an expander, then the sampler yields uncorrelated configuration with high probability after $t_f = O(\log|V|)$ steps ; the suggested t_f follows from the loose upper bound $|V| \leq m!$. A proof that $\mathcal{G}(L_{\max})$ is in fact an expander will depend on the specifics of $(\boldsymbol{d}, \boldsymbol{s}, L_{\max})$; however, we note that $\mathcal{G}(L_{\max})$ shares the two fundamental properties with all expanders for sufficiently large $L_{\max}$: It is connected and not bipartite.

simplex in the real system than in the randomized one, suggesting that some form of high-order triadic closure is at play [259]. The difference is, however, much more pronounced in the crime dataset ; this could be due to the fact that it describes a social system, whose structure tend to be heavily driven by triadic closure [103] (and potential high-order analogs).

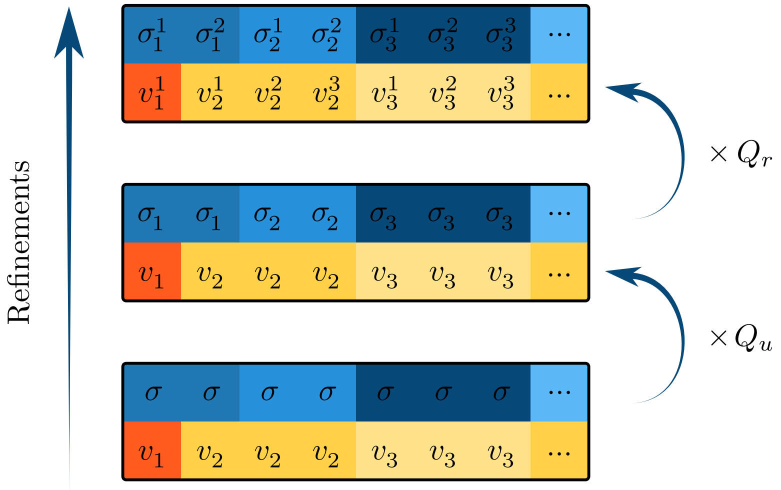
Finally, taking both distributions into account, we conclude that the shape of the pollinator dataset is completely determined by its local structure, while large–scale organizational principles influence the structure of the other datasets. This leads us to two final observations : One, care must be exerted in drawing conclusions about the shape of complex datasets—from the homology point of view there is nothing of note in the structure of the pollinator dataset. Two, some datasets—here the crime and diseasome datasets—are decidedly *not* random. This raises the question of just how much information must models account for, before they can capture such atypical Betti numbers. Would, for example, adding limited correlations among degrees be sufficient to capture the shape of most real datasets ? Or do we need to embrace growth models, with their sophisticated rules, and clustered local structure [22, 102, 242] ?

5.8 Perspectives

As it stands, the SCM already establishes the analysis of simplicial complexes on firmer statistical ground. The next step will be to clarify a number of important open questions, e.g., what is the true value of $L^*_{\max}$ for arbitrary simplicial complexes, and what is optimal choice of proposal distribution $\mathbb{P}$ (cf. Fig. 5.3) ?

Beyond these obvious questions, the connection between the SCM and the simple CM leads us to a series of natural problems not addressed in this paper. These include the problem of the *simpliciality* of arbitrary pairs of sequences (i.e., is there a simplicial complex which realizes a pair of sequences ?) [47], related to the problem of constructing initial conditions for the MCMC sampler, when no real system is available. We believe that the solution to such problems will require new insights, as the no–inclusion constraint appears to be a major obstacle to the application of classical methods developed for the analogous *graphicality* problem [92, 96].

In closing, we stress that all the above questions and challenges are of technical nature ; the model and sampler can already be applied to practical problems. This could lead to improvements in persistent homology (e.g. statistically sound filtrations of weighted complexes) or a formulation of community detection of simplicial complexes (via modularity [176]), and could provide a new glimpse into the emergence of homology and higher order structural properties in real complex systems.



Refinements
σ_1^1 σ_1^2 σ_2^1 σ_2^2 σ_3^1 σ_3^2 σ_3^3 $\cdots$
v_1^1 v_2^1 v_2^2 v_2^3 v_3^1 v_3^2 v_3^3 $\cdots$
$\times Q_r$
σ_1 σ_1 σ_2 σ_2 σ_3 σ_3 σ_3 $\cdots$
v_1 v_2 v_2 v_2 v_3 v_3 v_3 $\cdots$
$\times Q_u$
σ σ σ σ σ σ σ $\cdots$
v_1 v_2 v_2 v_2 v_3 v_3 v_3 $\cdots$

FIGURE 5.5 – Array representation of the sequences $(\boldsymbol{d},\boldsymbol{s}) = ([1,3,3,\ldots],[2,2,3\ldots])$ prior to shuffling. The order of appearance of nodes and facets in the sequences is assumed to correspond to their labels. (top) Stub-labeled ensemble (middle) fully-labeled ensemble (bottom) node-labeled ensemble, where the two size 2 facets are now indistinguishable. The (constant) size of the refinements is indicated on the left-hand size.

5.9 Supplementary Material

5.9.1 Labels do not matter

In this short paragraph, we prove that the stub-labeled SCM (where stubs in $B(K)$ are distinguishable) is a constant refinement of the fully-labeled SCM (where stubs are unlabeled, but facets are distinguishable), which is itself a constant refinement of the node-labeled SCM (where only nodes are distinguishable). By "x is constant refinement of y", we mean there is a constant number Q of elements in x associated to each element in y, such that $|x| = Q|y|$ and that a uniform distribution over x can be used to sample uniformly from y. This hierarchy of refinement implies that even when we are dealing with unlabeled simplicial complexes, we can think of facets and stubs as having labels, without introducing a bias in the sampler, calculations, etc. This practical property, also present in the configuration model defined on simple graphs [79], greatly simplifies the numerical implementation of sampling methods for the ensemble.

Proof

We harness the power of the array representation of simplicial complexes. For fixed degree and size sequences $(\boldsymbol{d},\boldsymbol{s})$ we create an array X of length $m = \sum_i d_i = \sum_i s_i$, where each facet appears s_i times, with distinct labels (i.e., $\sigma_i^1, \sigma_i^2, ..., \sigma_i^{s_i}$), and where each node appears d_i time (i.e., $v_i^1, v_i^2, ..., v_i^{d_i}$). An example is shown in Fig. 5.5. By generating a random permutation $\tilde{X}$ of, say, the row associated to nodes, we get a random matchings of the nodes and facets, i.e., a random bipartite graph $B(\tilde{X})$, which can then be interpreted as a simplicial complex

$K(\tilde{X})$ if $B(\tilde{X})$ is sequence-preserving. Therefore, the ensemble of all stub-labeled simplicial complexes can be represented with sequence-preserving arrays (array that correspond to sequence-preserving bipartite graphs).

To obtain refinement results, we will consider permutations of classes of elements in the arrays. First, notice that all stub-labeled arrays (top row in Fig. 5.5) associated to the same array X_r with labels removed from the stubs (i.e., v_i^j becomes v_i, see second row of Fig. 5.5) can be obtained by permuting the entries associated to each nodes / facet separately. There are

$$Q_r(\boldsymbol{d},\boldsymbol{s}) = \prod_{i=1}^{n} d_i! \prod_{i=1}^{f} s_i! \tag{5.4}$$

such permutations. Since $Q_r(\boldsymbol{d},\boldsymbol{s})$ only depends on the degree and size sequences, which are fixed for all elements of the SCM ensemble by definition, it follows that there is a constant number $Q_s(\boldsymbol{d},\boldsymbol{s})$ of fully-labeled arrays $\{X\}$ associated to each array X_r with stub labels removed. The statement obviously holds for the associated ensemble of bipartite graphs and simplicial complexes, which establishes that the stub-labeled SCM is a constant refinement of the fully-labeled SCM.

Now, if we also remove the labels from the *facets* and only distinguish them by size classes, we find by permuting the labels in the top row of X_r (middle row in Fig. 5.5) that there are

$$Q_u(\boldsymbol{d},\boldsymbol{s}) = \prod_{k=1}^{\max \boldsymbol{s}} q_k(\boldsymbol{s})! \tag{5.5}$$

arrays $\{X_r\}$ with labeled facets associated to each X_u, where $q_k(\boldsymbol{s})$ is the number of times that k appears in the sequence of integer $\boldsymbol{s}$ (its multiplicity). This immediately proves that the fully-labeled SCM is a constant refinement of the node-labeled SCM, but also that the stub-labeled SCM is a constant refinement of the node-labeled SCM, with scaling factor

$$Q(\boldsymbol{d},\boldsymbol{s}) = Q_r(\boldsymbol{d},\boldsymbol{s})Q_u(\boldsymbol{d},\boldsymbol{s}) = \prod_{i=1}^{n} d_i! \prod_{i=1}^{f} s_i! \prod_{k=1}^{\max \boldsymbol{s}} q_k(\boldsymbol{s})!\,, \tag{5.6}$$

by transitivity.

Although we do not consider it explicitly in the main text, note that an identical calculation connects the unlabeled SCM to the node-labeled SCM, with scaling factor $\prod_{k=1}^{\max \boldsymbol{s}} q_k(\boldsymbol{d})!$.

5.9.2 Detailed proof of Equation 5.2

Equation 5.2 (reproduced below) states that the fraction ϕ of sparse bipartite multigraphs with degree sequences $(\boldsymbol{d},\boldsymbol{s})$ that have no parallel edges tends to

$$\phi(\boldsymbol{d},\boldsymbol{s}) = e^{-\frac{1}{2}\left(\langle d^2\rangle/\langle d\rangle-1\right)\left(\langle s^2\rangle/\langle s\rangle-1\right)}\,,$$

where

$$\langle x^k \rangle = \frac{1}{|x|} \sum x_i^k$$

is the kth moment of the sequence x. This equation alone implies that sampling from the SCM using a stub-matching scheme is generally inefficient : A random stub-matching corresponds to a non-sequence-preserving configurations with probability bounded from below by

$$1 - \phi(\boldsymbol{d}, \boldsymbol{s}) \, .$$

Although ϕ is *not* a growing function of m (the number of stubs), it is large enough to cause practical problems, especially when the elements of $\boldsymbol{d}$ and $\boldsymbol{s}$ have a skewed empirical distribution. Indeed, to generate unbiased samples with stub-matchings, one must construct a random permutation of X and reject any non-sequence-preserving matching. Since permutations are independent, this implies that one will have to construct, on average, at least $\frac{1}{\phi} = e^{\frac{1}{2}\left(\langle d^2\rangle/\langle d\rangle - 1\right)\left(\langle s^2\rangle/\langle s\rangle - 1\right)}$ permutations before a success is encountered (there number of rejection follows a geometric distribution). The MCMC algorithm of the main text is much more efficient, with its $O(1)$ moves.

We note that the rejection problem is further compounded by the fact that many bipartite graphs with no parallel edge must also be rejected, due to included neighborhoods (see Fig. 5.2 of the main text). But since Eq. (5.2) is enough to establish that stub-matching is impractical, and since bounding or calculating the latter probability is much more involved, we will be content with the result proved below. An approach similar to that of Ref. [153, Chapter 9] could be used to refine the bound $1 - \phi$.

Proof

Let us consider all permutations of a fixed array X constructed with a fixed pairs of sequences $(\boldsymbol{d}, \boldsymbol{s})$, i.e., all stub-matchings, in the array representation. We introduce the following notation for events

$$\begin{aligned} E_{ijk} : &\quad (v_i, \sigma_j) \text{ are aligned } k \text{ times in X,} \\ \mu_{ij} : &\quad (v_i, \sigma_j) \text{ are aligned } k \geq 2 \text{ times in X.} \end{aligned}$$

and investigate a uniform distribution over this set. From the definition of the distribution it immediately follows that the number of edges k between node v_i and facet σ_j in X is drawn from the hypergeometric distribution, i.e.,

$$\Pr[E_{ijk}] = \frac{\binom{m-s_j}{d_i-k}\binom{s_j}{k}}{\binom{m}{d_i}} \, . \tag{5.7}$$

The probability that (v_i,σ_j) are aligned more than one time (i.e., the probability that there is a parallel edge involving v_i and σ_j) is therefore

$$\begin{aligned}\Pr[\mu_{ij}] &= 1 - \frac{\binom{m-s_j}{d_i}}{\binom{m}{d_i}} - \frac{\binom{m-s_j}{d_i-1}s_j}{\binom{m}{d_i}}\\ &= 1 - \left[1 - \frac{d_i s_j}{m} + \frac{d_i s_j(d_i-1)(s_j-1)}{2m^2} + O(m^{-3})\right]\\ &\qquad\qquad - \left[\frac{d_i s_j}{m} - \frac{d_i s_j(d_i-1)(s_j-1)}{m^2} + O(m^{-3})\right]\\ &= \frac{1}{2}\frac{d_i(d_i-1)s_j(s_j-1)}{m^2} + O(m^{-3})\,, \end{aligned} \tag{5.8}$$

where the series at the second equalities are computed at $m=\infty$, with d_i,s_j constant.

The events $\{\mu_{ij}\}$ are not strictly independent from one another, but $\Pr[E_{ijk}|E_{abk'}] \to \Pr[E_{ijk}]$ in the sparse limits ($m\to\infty$ with bounded first moments $\langle d\rangle$ and $\langle s\rangle$), Thus, we can treat the events as *asymptotically* independent. This allows us to write

$$\Pr[\text{No multiedges}] \approx \prod_{i,j}(1 - Pr[\mu_{ij}]) = \prod_{i,j}\left[1 - \frac{1}{2}\frac{d_i(d_i-1)s_j(s_j-1)}{m^2}\right]\,. \tag{5.9}$$

To show that Eq. (5.9) predicts that the fraction of matchings with no multiedges depends exponentially on the ratio of moments $(\langle s^2\rangle/\langle s\rangle, \langle d^2\rangle/\langle d\rangle)$, we rewrite the product over facets and vertices as a product over degrees and sizes, and then use the definition of the exponential function. Recalling that $q_k(\boldsymbol{x})$ is the multiplicity of k in sequence $\boldsymbol{x}$, that n is the number of nodes and f the number of facets, this approach yields Eq. (5.2) directly :

$$\begin{aligned}\Pr[\text{No multiedges}] &\approx \prod_{d,s}\left[1 - \frac{1}{2}\frac{d(d-1)s(s-1)}{nf\langle s\rangle\langle d\rangle}\right]^{nfq_d(\boldsymbol{d})q_s(\boldsymbol{s})/nf}\,,\\ &= \prod_{d,s}\exp\left[-\frac{1}{2}\frac{d(d-1)s(s-1)}{\langle s\rangle\langle d\rangle}\frac{q_d(\boldsymbol{d})q_s(\boldsymbol{s})}{nf}\right]\,,\\ &= \exp\left[-\frac{1}{2}\sum_{d,s}\frac{d(d-1)s(s-1)}{\langle s\rangle\langle d\rangle}\frac{q_d(\boldsymbol{d})q_s(\boldsymbol{s})}{nf}\right]\,,\\ &= \exp\left[-\frac{1}{2}\frac{(\langle d^2\rangle - \langle d\rangle)(\langle s^2\rangle - \langle s\rangle)}{\langle s\rangle\langle d\rangle}\right]\,, \end{aligned} \tag{5.10}$$

validated in Fig. 5.6 .

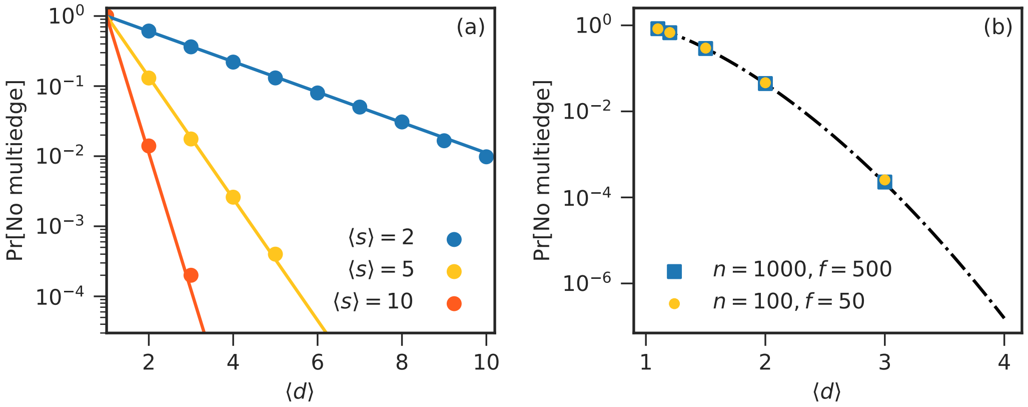


FIGURE 5.6 – Validation of Eq. (5.2). In both figures, symbols are obtained by averaging over 5 000 random matchings, while Eq. (5.2) is shown with solid line(s). (a) Fraction of random regular stub-matchings that have no multiedges, as a function of the degree of all nodes $\langle d \rangle$. These matchings are constructed with $n = 1000$ nodes and $f = \langle d \rangle n / \langle s \rangle$ facets of fixed size $\langle s \rangle = [2, 5, 10]$. All quantities are fixed upon choosing a fixed degree and facet size, since there must be a matching number of stubs of each types, i.e., $\langle d \rangle / \langle s \rangle = f/n$. (b) Fraction of zero-truncated Poisson (ZTP) sub matchings that have no multiedges, as a function of the expected degree $\langle d \rangle$. These matchings are constructed by drawing random sequences $(\boldsymbol{d}, \boldsymbol{s})$ from two ZTP distributions. The ZTP distribution is a Poisson distributions with no weight on zero, and with the PMF $p_k(\lambda) = \lambda^k / [(e^\lambda - 1)k!]$ for $k > 1$ Again, the condition $\langle d \rangle / \langle s \rangle = f/n$ fixes the expectations of the elements of $(\boldsymbol{d}, \boldsymbol{s})$, this time upon choosing n and f. Assuming that sequences are typical with high probability, we can simply replace $\langle x^k \rangle$ by $\mathbb{E}[x^k]$ in Eq. (5.2), where $\mathbb{E}[x^k]$ is the expectation of a *distribution*, instead of a moment of a fixed series. The good agreement of theoretical and simulated results show that Eq. (5.2) works well, even for small simplicial complexes constructed from typical random sequences $(\boldsymbol{d}, \boldsymbol{s})$.

Conclusion

La science des réseaux est récemment arrivée à maturité. Bien que la direction à suivre ne soit pas encore complètement claire, une chose certaine est que cette route passera par une approche plus rigoureuse, après le foisonnement de modèles et d'explorations des premières années.

Dans cette thèse, nous avons montré à l'aide de problèmes concrets que l'inférence statistique devra faire partie—et fait même déjà partie—de cette nouvelle vague. Du point de vue *structurel*, nous avons conclu que le modèle stochastique par blocs (SBM) est une clé de lecture nécessaire à la bonne compréhension de la structure de tout réseau (partie I), d'une part parce qu'il est universel (Chap. 2) et d'autre part parce qu'il est cohérent (Chap. 3). Pour ce faire, nous avons adopté une approche statistique rigoureuse, plutôt que l'approche algorithmique *ad hoc* plus traditionnelle. On a ainsi pu illustrer la puissance du paradigme de l'inférence réseau. Du point de vue *temporel*, nous avons montré qu'on peut comprendre la reconstruction du passé comme un problème d'estimation bayésien (partie II). Nous avons utilisé le modèle de l'attachement préférentiel comme modèle génératif (Chap. 4), ce qui nous a permis de récupérer l'histoire de réseaux autant artificiels que réels. Finalement, au-delà des réseaux, nous avons démontré qu'on peut combiner une généralisation des réseaux, homologie, et test d'hypothèses pour analyser la structure des systèmes complexes et quantifier leur niveau d'organisation (Chap. 5).

Au fil des chapitres, nous avons identifié et laissé en plan plusieurs problèmes : des généralisations trop éloignées de la thématique principale de la thèse, des problèmes plus techniques, ou encore des suites naturelles escamotées par manque de temps. Les plus intéressants d'entre eux sont probablement :

- ◇ **Chapitres 2 et 3** (SBM) : L'extension de l'analyse de cohérence au cas infini et l'identification d'un critère permettant de sélectionner la bonne famille de courbes. Avec ces deux résultats, on pourrait obtenir une preuve plus simple de résultats déjà connus pour le régime infini [1].
- ◇ **Chapitre 4** (Archéologie réseau) :
  - ○ Le développement d'une méthode d'inférence efficace *et* exacte (ou quasi exacte), par exemple par passation de message dynamique [143]. Pour l'instant nos mé-

thodes exactes sont coûteuses, et l'erreur de nos méthodes efficaces n'est pas bornée.

  - ○ Une meilleure caractérisation de la transition de phase, e.g., l'identification de sa position exacte (nous avons conjecturé que $\gamma_c = 3/2$).
  - ○ L'extension à des modèles autres que l'attachement préférentiel, par exemple les modèles avec *fitness* [21].

- ◇ **Chapitre 5** (SCM) :
  - ○ Une preuve que la chaîne de Markov est connectée pour un $L^*_{\max}$ quelconque.
  - ○ Un test constructif analogue à celui de Havel et Hakimi [92, 96], permettant de garantir qu'une paire de séquences $(\boldsymbol{d}, \boldsymbol{s})$ est associée à au moins un complexe simplicial.
  - ○ Une formule fermée ou une méthode de calcul efficace pour la taille espérée des groupes d'homologie, pour des séquences $(\boldsymbol{d}, \boldsymbol{s})$ quelconques.

Malgré cette foule de questions ouvertes, on peut affirmer, avec certitude, que les travaux de cette thèse ont déjà des applications immédiates. Par exemple, le SBM peut désormais être utilisé pour générer des méthodes de détection de régularité mésoscopique sur mesure, de façon quasi automatique [216]. Ces méthodes sont, de surcroît, accompagnées de garanties de cohérence—même en taille finie—grâce à l'analyse du chapitre 3. Pour ce qui a trait à l'inférence temporelle (Chap. 4), notre approche originale de la reconstruction du passé des réseaux pourra faire office de base pour le développement d'un nouveau champ de recherche, soit celui de l'archéologie réseau. Les méthodes du chapitre 5, quant à elles, établissent une bonne partie des fondements nécessaires à une théorie des complexes simpliciaux aléatoires, inspirée par la théorie réseau analogue, introduite il y a près de vingt ans [178].

Si on a mis beaucoup l'accent sur un traitement statistique rigoureux, c'est que la science des réseaux pourra en bénéficier grandement. Il ne s'agit toutefois pas de la seule avenue qui vaudra la peine d'être explorée plus en profondeur dans les années à venir. Ainsi, les travaux de cette thèse—surtout ceux de la deuxième partie et l'épilogue—mettent aussi la table pour une science de la structure des systèmes complexes, au-delà des réseaux. En s'attardant à l'histoire des réseaux (Chap. 4), à la méta-information (Chap. 2) et aux interactions à plusieurs corps (Chap. 5), on a implicitement utilisé de l'information normalement laissée de côté par les méthodes réseaux plus naïves. Il reste à voir si ces approches, dites d'ordre supérieur, seront appelées à jouer un rôle plus important dans le futur.

Besser organisiert,
wäre ich gefährlich

Proverbe serbo-croate

Bibliographie

[1] E. ABBE, *Community detection and stochastic block models : Recent developments*, J. Mach. Learn. Res., 18 (2018), p. 1–86.

[2] E. ABBE, A. S. BANDEIRA ET G. HALL, *Exact recovery in the stochastic block model*, IEEE Transactions on Information Theory, 62 (2016), p. 471–487.

[3] E. ABBE ET C. SANDON, *Community detection in general stochastic block models : Fundamental limits and efficient algorithms for recovery*, dans Proceedings of the 2015 IEEE 56th Annual Symposium on Foundations of Computer Science, IEEE, Washington DC, 2015, p. 670–688.

[4] L. A. ADAMIC ET B. A. HUBERMAN, *Power-law distribution of the World Wide Web*, Science, 287 (2000).

[5] R. ALBERT ET A.-L. BARABÁSI, *Topology of evolving networks : local events and universality*, Phys. Rev. Lett., 85 (2000), p 5234.

[6] ——, *Statistical mechanics of complex networks*, Rev. Mod. Phys., 74 (2002), p 47.

[7] A. ALLARD, *Percolation sur graphes aléatoires-modélisation et description analytique*, Thèse doctorat, Université Laval, 2014.

[8] A. ALLARD, B. M. ALTHOUSE, S. V. SCARPINO ET L. HÉBERT-DUFRESNE, *Asymmetric percolation drives a double transition in sexual contact networks*, Proc. Natl. Acad. Sci. U.S.A., 114 (2017), p. 8969–8973.

[9] A. ALLARD, L. HÉBERT-DUFRESNE, J.-G. YOUNG ET L. J. DUBÉ, *General and exact approach to percolation on random graphs*, Phys. Rev. E, (2015).

[10] A. ALLARD, P.-A. NOËL, L. J. DUBÉ ET B. POURBOHLOUL, *Heterogeneous bond percolation on multitype networks with an application to epidemic dynamics*, Phys. Rev. E, 79 (2009), p 036113.

[11] C. ANDRIEU, N. DE FREITAS, A. DOUCET ET M. I. JORDAN, *An introduction to MCMC for machine learning*, Mach. Learn., 50 (2003), p. 5–43.

[12] A. ARENAS, J. DUCH, A. FERNÁNDEZ ET S. GÓMEZ, *Size reduction of complex networks preserving modularity*, New J. Phys., 9 (2007), p 176.

[13] B. BALL, B. KARRER ET M. E. J. NEWMAN, *Efficient and principled method for detecting*

communities in networks, Phys. Rev. E, 84 (2011), p 036103.

[14] J. BANKS, C. MOORE, J. NEEMAN ET P. NETRAPALLI, *Information-theoretic thresholds for community detection in sparse networks*, dans Proceedings of the 29th Annual Conference on Learning Theory, 2016, p. 383–416.

[15] A.-L. BARABÁSI, *The network takeover*, Nat. Phys., 8 (2011), p 14.

[16] A.-L. BARABÁSI ET R. ALBERT, *Emergence of scaling in random networks*, Science, 286 (1999), p. 509–512.

[17] M. J. BARBER ET J. W. CLARK, *Detecting network communities by propagating labels under constraints*, Phys. Rev. E, 80 (2009), p 026129.

[18] V. BATAGELJ ET M. ZAVERSNIK, *An $O(m)$ algorithm for cores decomposition of networks*, arXiv :cs/0310049, (2003).

[19] E. A. BENDER, *The asymptotic number of non-negative integer matrices with given row and column sums*, Discrete Math., 10 (1974), p. 217–223.

[20] E. A. BENDER ET E. R. CANFIELD, *The asymptotic number of labeled graphs with given degree sequences*, J. Combin. Theo. A, 24 (1978), p. 296–307.

[21] G. BIANCONI ET A.-L. BARABÁSI, *Bose-Einstein condensation in complex networks*, Phys. Rev. Lett., 86 (2001), p 5632.

[22] G. BIANCONI ET C. RAHMEDE, *Network geometry with flavor : From complexity to quantum geometry*, Phys. Rev. E, 93 (2016), p 032315.

[23] P. J. BICKEL ET A. CHEN, *A nonparametric view of network models and Newman–Girvan and other modularities*, Proc. Natl. Acad. Sci. U.S.A., 106 (2009), p. 21068–21073.

[24] J. BLITZSTEIN ET P. DIACONIS, *A sequential importance sampling algorithm for generating random graphs with prescribed degrees*, Internet Math., 6 (2011), p. 489–522.

[25] V. D. BLONDEL, J.-L. GUILLAUME, R. LAMBIOTTE ET E. LEFEBVRE, *Fast unfolding of communities in large networks*, J. Stat. Mech. Theor. Exp., 2008 (2008), p P10008.

[26] B. BOLLOBÁS, *A probabilistic proof of an asymptotic formula for the number of labelled regular graphs*, Eur. J. Combin., 1 (1980), p. 311–316.

[27] ——, *Modern Graph Theory*, Springer, New York, 1998.

[28] S. P. BORGATTI ET M. G. EVERETT, *Models of core/periphery structures*, Soc. Networks, 21 (2000), p. 375–395.

[29] C. P. BORRAS, L. HERNANDEZ ET Y. MORENO, *Breaking the spell of nestedness*, arXiv :1711.03134, (2017).

[30] S. BOYD ET L. VANDENBERGHE, *Convex Optimization*, Cambridge University Press, Cambridge, 2004.

[31] U. BRANDES, D. DELLING, M. GAERTLER, R. GÖRKE, M. HOEFER, Z. NIKOLOSKI ET D. WAGNER, *Maximizing modularity is hard*, arXiv :0608255, (2006).

[32] S. BUBECK, L. DEVROYE ET G. LUGOSI, *Finding Adam in random growing trees*, Random Struct. Algor., 50 (2017), p. 158–172.

[33] S. BUBECK, R. ELDAN, E. MOSSEL, M. Z. RÁCZ *et al.*, *From trees to seeds : on the inference of the seed from large trees in the uniform attachment model*, Bernoulli, 23 (2017), p. 2887–2916.

[34] A. CAIMO ET N. FRIEL, *Bayesian inference for exponential random graph models*, Soc. Networks, 33 (2011), p. 41–55.

[35] D. S. CALLAWAY, M. E. J. NEWMAN, S. H. STROGATZ ET D. J. WATTS, *Network robustness and fragility : percolation on random graphs*, Phys. Rev. Lett., 85 (2000), p 5468.

[36] G. CASELLA ET R. L. BERGER, *Statistical Inference*, Duxbury, Pacific Grove, 1ere éd., 2002.

[37] C. CASTELLANO ET R. PASTOR-SATORRAS, *Relating topological determinants of complex networks to their spectral properties : Structural and dynamical effects*, Phys. Rev. X, 7 (2017), p 041024.

[38] J. M. CHAN, G. CARLSSON ET R. RABADAN, *Topology of viral evolution*, Proc. Natl. Acad. Sci. U.S.A., 110 (2013), p. 18566–18571.

[39] Z. CHANG, H.-M. CHENG, C. YAN, X. YIN ET Z.-Y. ZHANG, *On approximate equivalence of modularity, D and non-negative matrix factorization*, arXiv :1801.03618, (2018).

[40] F. CHUNG ET L. LU, *The average distances in random graphs with given expected degrees*, Proc. Natl. Acad. Sci. U.S.A., 99 (2002), p. 15879–15882.

[41] A. CLAUSET, C. MOORE ET M. E. NEWMAN, *Hierarchical structure and the prediction of missing links in networks*, Nature, 453 (2008), p. 98–101.

[42] A. CLAUSET, C. R. SHALIZI ET M. E. NEWMAN, *Power-law distributions in empirical data*, SIAM Rev., 51 (2009), p. 661–703.

[43] A. CONDON ET R. M. KARP, *Algorithms for graph partitioning on the planted partition model*, Rand. Struct. Alg., 18 (2001), p. 116–140.

[44] E. F. CONNOR ET D. SIMBERLOFF, *The assembly of species communities : chance or competition ?*, Ecology, 60 (1979), p. 1132–1140.

[45] T. COOLEN, A. ANNIBALE ET E. ROBERTS, *Generating Random Networks and Graphs*, Oxford University Press, Oxford, 2017.

[46] A. COSTA ET M. FARBER, *Random Simplicial Complexes*, dans Configuration Spaces, Springer, Berlin, 2016, p. 129–153.

[47] O. T. COURTNEY ET G. BIANCONI, *Generalized network structures : The configuration model and the canonical ensemble of simplicial complexes*, Phys. Rev. E, 93 (2016), p 062311.

[48] T. M. COVER ET J. A. THOMAS, *Elements of Information Theory*, John Wiley & Sons, New York, 2012.

[49] H. S. M. COXETER, *Regular Polytopes*, Courier Corporation, New York, 1973.

[50] P. CRESCENZI ET V. KANN, *Approximation on the Web : A Compendium of NP Optimization Problems*, 1997.

[51] C. CURTO ET V. ITSKOV, *Cell groups reveal structure of stimulus space*, PLoS Comput. Biol., 4 (2008), p e1000205.

[52] Y. DABAGHIAN, F. MÉMOLI, L. FRANK ET G. CARLSSON, *A topological paradigm for hippocampal spatial map formation using persistent homology*, PLoS Comput. Biol., 8 (2012), p e1002581.

[53] L. DANON, A. DIAZ-GUILERA, J. DUCH ET A. ARENAS, *Comparing community structure identification*, J. Stat. Mech. Theor. Exp., 2005 (2005), p P09008.

[54] C. DE BACCO, D. B. LARREMORE ET C. MOORE, *A physical model for efficient ranking in networks*, arXiv :1709.09002, (2017).

[55] A. DECELLE, F. KRZAKALA, C. MOORE ET L. ZDEBOROVÁ, *Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications*, Phys. Rev. E, 84 (2011), p 066106.

[56] ——, *Inference and phase transitions in the detection of modules in sparse networks*, Phys. Rev. Lett., 107 (2011), p 065701.

[57] S. DECKER, C. W. KOHFELD, R. ROSENFELD ET J. SPRAGUE, *St. Louis Homicide Project : Local Responses to a National Problem*, University of Missouri, St. Louis, 1991.

[58] P. DIACONIS ET D. YLVISAKER, *Conjugate priors for exponential families*, Ann. Stat., 7 (1979), p. 269–281.

[59] P. DIAO, D. GUILLOT, A. KHARE ET B. RAJARATNAM, *Model-free consistency of graph partitioning*, arXiv :1608.03860, (2016).

[60] P. S. DODDS, D. R. DEWHURST, F. F. HAZLEHURST, C. M. VAN OORT, L. MITCHELL, A. J. REAGAN, J. R. WILLIAMS ET C. M. DANFORTH, *Simon's fundamental rich-get-richer model entails a dominant first-mover advantage*, Phys. Rev. E, 95 (2017), p 052301.

[61] S. N. DOROGOVTSEV ET J. F. MENDES, *Evolution of networks*, Adv. Phys., 51 (2002), p. 1079–1187.

[62] S. N. DOROGOVTSEV ET J. F. F. MENDES, *Scaling behaviour of developing and decaying networks*, Europhys. Lett., 52 (2000), p 33.

[63] S. N. DOROGOVTSEV, J. F. F. MENDES ET A. N. SAMUKHIN, *Structure of growing networks with preferential linking*, Phys. Rev. Lett., 85 (2000), p 4633.

[64] M. DRMOTA, *Random trees : An Interplay Between Combinatorics and Probability*, Springer, New York, 2009.

[65] R. DUNBAR, *Neocortex size as a constraint on group size in primates*, J. Hum. Evol., 22 (1992), p. 469–493.

[66] J. DUTKOWSKI ET J. TIURYN, *Identification of functional modules from conserved ancestral protein–protein interactions*, Bioinformatics, 23 (2007), p. i149–i158.

[67] S. P. ELLIS ET A. KLEIN, *Describing high-order statistical dependence using "concurrence topology" with application to functional mri brain data*, Homology, Homotopy Appl., 16 (2014), p. 245–264.

[68] P. ERDŐS ET A. RÉNYI, *On random graphs I*, Publicationes Mathematicae, 6 (1959), p. 290–297.

[69] ——, *On the evolution of random graphs*, Publ. Math. Inst. Hungar. Acad. Sci, 5 (1960), p. 17–61.

[70] B. H. ERICKSON, *Some problems of inference from chain data*, Sociol. Methodol., 10 (1979), p. 276–302.

[71] P. EXPERT, T. S. EVANS, V. D. BLONDEL ET R. LAMBIOTTE, *Uncovering space-independent communities in spatial networks*, Proc. Natl Acad. of Sci., 108 (2011), p. 7663–7668.

[72] S. E. FIENBERG, M. M. MEYER ET S. S. WASSERMAN, *Statistical analysis of multiple sociometric relations*, J. Am. Stat. Assoc., 80 (1985), p. 51–67.

[73] S. E. FIENBERG ET S. S. WASSERMAN, *Categorical data analysis of single sociometric relations*, Sociol. Methodol., 12 (1981), p. 156–192.

[74] R. FISCHER, J. C. LEITAO, T. P. PEIXOTO ET E. G. ALTMANN, *Sampling motif-constrained ensembles of networks*, Phys. Rev. Lett., 115 (2015), p 188701.

[75] J. FLANNICK, A. NOVAK, B. S. SRINIVASAN, H. H. MCADAMS ET S. BATZOGLOU, *Graemlin : general and robust alignment of multiple large interaction networks*, Genome Res., 16 (2006), p. 1169–1181.

[76] S. FORTUNATO, *Community detection in graphs*, Phys. Rep., 486 (2010), p. 75–174.

[77] S. FORTUNATO ET M. BARTHELEMY, *Resolution limit in community detection*, Proc. Natl. Acad. Sci. U.S.A., 104 (2007), p. 36–41.

[78] S. FORTUNATO ET D. HRIC, *Community detection in networks : A user guide*, Phys. Rep., 659 (2016), p. 1–44.

[79] B. K. FOSDICK, D. B. LARREMORE, J. NISHIMURA ET J. UGANDER, *Configuring random graph models with fixed degree sequences*, arXiv :1608.00607, (2018).

[80] O. FRANK ET D. STRAUSS, *Markov graphs*, J. Am. Stat. Assoc., 81 (1986), p. 832–842.

[81] T. GAUDELET, N. MALOD-DOGNIN ET N. PRZULJ, *Higher order molecular organisation as a source of biological function*, arXiv :1804.05003, (2018).

[82] A. GHASEMIAN, H. HOSSEINMARDI ET A. CLAUSET, *Evaluating overfit and underfit in models of network community structure*, arXiv :1802.10582, (2018).

[83] T. A. GIBSON ET D. S. GOLDBERG, *Reverse engineering the evolution of protein interaction*

networks, dans Biocomputing, World Scientific, 2009, p. 190–202.

[84] E. N. GILBERT, *Random graphs*, Ann. Math. Stat., 30 (1959), p. 1141–1144.

[85] C. GIUSTI, R. GHRIST ET D. S. BASSETT, *Two's company, three (or more) is a simplex*, J. Comput. Neurosci., 41 (2016), p. 1–14.

[86] K.-I. GOH, M. E. CUSICK, D. VALLE, B. CHILDS, M. VIDAL ET A.-L. BARABÁSI, *The human disease network*, Proc. Natl. Acad. Sci. U.S.A., 104 (2007), p. 8685–8690.

[87] V. GÓMEZ, H. J. KAPPEN ET A. KALTENBRUNNER, *Modeling the structure and evolution of discussion cascades*, dans Proceedings of the 22nd ACM conference on Hypertext and hypermedia, ACM, 2011, p. 181–190.

[88] M. S. GRANOVETTER, *The strength of weak ties*, dans Soc. Networks, Elsevier, 1977, p. 347–367.

[89] T. GROSS, *Complex networks : Don't call in sick*, Nat. Phys., 12 (2016), p. 995–996.

[90] A. HAGBERG, P. SWART ET D. S CHULT, *Exploring network structure, dynamics, and function using NetworkX*, Los Alamos National Laboratory (LANL), 2008.

[91] B. HAJEK ET S. SANKAGIRI, *Recovering a hidden community in a preferential attachment graph*, arXiv :1801.06818, (2018).

[92] S. L. HAKIMI, *On realizability of a set of integers as degrees of the vertices of a linear graph*, J. Soc. Ind. Appl. Math., 10 (1962), p. 496–506.

[93] M. S. HANDCOCK ET K. J. GILE, *Comment : on the concept of snowball sampling*, Sociol. Methodol., 41 (2011), p. 367–371.

[94] F. HARARY ET R. Z. NORMAN, *Some properties of line digraphs*, Rendiconti del Circolo Matematico di Palermo, 9 (1960), p. 161–168.

[95] A. HATCHER, *Algebraic Topology*, Cambridge University Press, Cambridge, 2000.

[96] V. HAVEL, *A remark on the existence of finite graphs*, Casopis Pest. Mat., 80 (1955), p. 477–480.

[97] L. HÉBERT-DUFRESNE, A. ALLARD, V. MARCEAU, P.-A. NOËL ET L. J. DUBÉ, *Structural preferential attachment : network organization beyond the link*, Phys. Rev. Lett., 107 (2011), p 158702.

[98] L. HÉBERT-DUFRESNE, A. ALLARD, V. MARCEAU, P.-A. NOËL ET L. J. DUBÉ, *Structural preferential attachment : stochastic process for the growth of scale-free, modular, and self-similar systems*, Phys. Rev. E, 85 (2012), p 026108.

[99] L. HÉBERT-DUFRESNE, A. ALLARD, P.-A. NOËL, J.-G. YOUNG ET E. LIBBY, *Strategic tradeoffs in competitor dynamics on adaptive networks*, arXiv :1607.04632, (2016).

[100] L. HÉBERT-DUFRESNE, A. ALLARD, J.-G. YOUNG ET L. J. DUBÉ, *Constrained growth of complex scale-independent systems*, Phys. Rev. E, 93 (2016), p 032304.

[101] L. HÉBERT-DUFRESNE, J. A. GROCHOW ET A. ALLARD, *Multi-scale structure and topo-*

logical anomaly detection via a new network statistic : the onion decomposition, Sci. Rep., 6 (2016).

[102] L. Hébert-Dufresne, E. Laurence, A. Allard, J.-G. Young et L. J. Dubé, *Complex networks as an emerging property of hierarchical preferential attachment*, Phys. Rev. E, 92 (2015), p 062809.

[103] C. A. Hidalgo, *Disconnected, fragmented, or united ? a trans-disciplinary review of network science*, Appl. Netw. Sci., 1 (2016), p 6.

[104] Y. Hiraoka, T. Nakamura, A. Hirata, E. G. Escolar, K. Matsue et Y. Nishiura, *Hierarchical structures of amorphous solids characterized by persistent homology*, Proc. Natl. Acad. Sci. U.S.A., 113 (2016), p. 7035–7040.

[105] P. W. Holland, K. B. Laskey et S. Leinhardt, *Stochastic blockmodels : First steps*, Soc. Networks, 5 (1983), p. 109–137.

[106] P. W. Holland et S. Leinhardt, *An exponential family of probability distributions for directed graphs*, J. Am. Stat. Assoc., 76 (1981), p. 33–50.

[107] D. Horak, S. Maletić et M. Rajković, *Persistent homology of complex networks*, J. Stat. Mech. Theor. Exp., 2009 (2009), p P03034.

[108] D. Hric, T. P. Peixoto et S. Fortunato, *Network structure, metadata, and the prediction of missing nodes and annotations*, Phys. Rev. X, 6 (2016), p 031038.

[109] D. Hu, P. Ronhovde et Z. Nussinov, *Phase transitions in random potts systems and the community detection problem : spin-glass type and dynamic perspectives*, Philos. Mag., 92 (2012), p. 406–445.

[110] A. Z. Jacobs et A. Clauset, *A unified view of generative models for networks : models, methods, opportunities, and challenges*, arXiv :1411.4070, (2014).

[111] A. Jasra, A. Persing, A. Beskos, K. Heine et M. De Iorio, *Bayesian inference for duplication–mutation with complementarity network models*, J. Comput. Biol., 22 (2015), p. 1025–1033.

[112] E. T. Jaynes, *Information theory and statistical mechanics*, Phys. Rev., 106 (1957), p 620.

[113] H. Jeong, Z. Néda et A.-L. Barabási, *Measuring preferential attachment in evolving networks*, Europhys. Lett., 61 (2003), p 567.

[114] M. Jerrum et G. B. Sorkin, *The Metropolis algorithm for graph bisection*, Discrete Appl. Math, 82 (1998), p. 155–175.

[115] M. Kahle, *Topology of random simplicial complexes : a survey*, AMS Contemp. Math., 620 (2014), p. 201–222.

[116] B. Karrer et M. E. J. Newman, *Stochastic blockmodels and community structure in networks*, Phys. Rev. E, 83 (2011), p 016107.

[117] M. Kato, T. Kakutani, T. Inoue et T. Itino, *Insect–flower relationship in the primary*

beech forest of Ashu, Contr. Biol. Lab. Kyoto Univ., 27 (1990), p. 309–375.

[118] T. KAWAMOTO, *Algorithmic infeasibility of community detection in higher-order networks*, arXiv :1710.08816, (2017).

[119] ——, *Algorithmic detectability threshold of the stochastic block model*, Phys. Rev. E, 97 (2018), p 032301.

[120] T. KAWAMOTO ET Y. KABASHIMA, *Detectability of the spectral method for sparse graph partitioning*, Europhys. Lett., 112 (2015), p 40007.

[121] ——, *Detectability thresholds of general modular graphs*, Phys. Rev. E, 95 (2017), p 012304.

[122] ——, *Comparative analysis on the selection of number of clusters in community detection*, Phys. Rev. E, 97 (2018), p 022315.

[123] B. W. KERNIGHAN ET S. LIN, *An efficient heuristic procedure for partitioning graphs*, Bell Syst. Tech. J, 49 (1970), p. 291–307.

[124] S. KIRKPATRICK, C. D. GELATT ET M. P. VECCHI, *Optimization by simulated annealing*, Science, 220 (1983), p. 671–680.

[125] S. KOJAKU ET N. MASUDA, *Finding multiple core-periphery pairs in networks*, Phys. Rev. E, 96 (2017), p 052313.

[126] P. L. KRAPIVSKY ET S. REDNER, *Organization of growing random networks*, Phys. Rev. E, 63 (2001), p 066123.

[127] P. L. KRAPIVSKY ET S. REDNER, *Statistics of changes in lead node in connectivity-driven networks*, Phys. Rev. Lett., 89 (2002), p 258703.

[128] P. L. KRAPIVSKY, S. REDNER ET F. LEYVRAZ, *Connectivity of growing random networks*, Phys. Rev. Lett., 85 (2000), p 4629.

[129] P. L. KRAPIVSKY, G. J. RODGERS ET S. REDNER, *Degree distributions of growing networks*, Phys. Rev. Lett., 86 (2001), p 5401.

[130] F. KRZAKALA, C. MOORE, E. MOSSEL, J. NEEMAN, A. SLY, L. ZDEBOROVÁ ET P. ZHANG, *Spectral redemption in clustering sparse networks*, Proc. Natl. Acad. Sci. U.S.A., 110 (2013), p. 20935–20940.

[131] F. KRZAKALA ET L. ZDEBOROVÁ, *Hiding quiet solutions in random constraint satisfaction problems*, Phys. Rev. Lett., 102 (2009), p 238701.

[132] T. O. KVÅLSETH, *Entropy and correlation : Some comments*, IEEE Trans. Syst., Man, Cybern., 17 (1987), p. 517–519.

[133] D. B. LARREMORE, A. CLAUSET ET A. Z. JACOBS, *Efficiently inferring community structure in bipartite networks*, Phys. Rev. E, 90 (2014), p 012805.

[134] E. LAURENCE, J.-G. YOUNG, S. MELNIK ET L. J. DUBÉ, *Exact analytical solution of irreversible binary dynamics on networks*, Phys. Rev. E, 97 (2018).

[135] S. H. LEE, P.-J. KIM ET H. JEONG, *Statistical properties of sampled networks*, Phys. Rev.

E, 73 (2006), p 016102.

[136] T. LESIEUR, F. KRZAKALA ET L. ZDEBOROVÁ, *MMSE of probabilistic low-rank matrix estimation : Universality with respect to the output channel*, dans Proceedings of the 2015 53rd Annual Allerton Conference on Communication, IEEE, 2015, p. 680–687.

[137] J. LESKOVEC, L. BACKSTROM, R. KUMAR ET A. TOMKINS, *Microscopic evolution of social networks*, dans Proceedings of the 14th ACM SIGKDD International Conference on Knowledge discovery and Data Mining, ACM, 2008, p. 462–470.

[138] S. LI, K. CHOI, T. WU ET L. ZHANG, *Reconstruction of network evolutionary history from extant network topology and duplication history*, dans Proceedings of the 2012 International Symposium on Bioinformatics Research and Applications, Springer, 2012, p. 165–176.

[139] S. LI, K. P. CHOI, T. WU ET L. ZHANG, *Maximum likelihood inference of the evolutionary history of a ppi network from the duplication history of its proteins*, IEEE T. Comput. Biol. Bioinformat., 10 (2013), p. 1412–1421.

[140] S. W. LINDERMAN, G. E. MENA, H. COOPER, L. PANINSKI ET J. P. CUNNINGHAM, *Reparameterizing the Birkhoff polytope for variational permutation inference*, arXiv :1710.09508, (2017).

[141] J. S. LIU ET R. CHEN, *Sequential Monte Carlo methods for dynamic systems*, J. Am. Stat. Assoc., 93 (1998), p. 1032–1044.

[142] Y.-Y. LIU ET A.-L. BARABÁSI, *Control principles of complex systems*, Rev. Mod. Phys., 88 (2016), p 035006.

[143] A. Y. LOKHOV, M. MÉZARD, H. OHTA ET L. ZDEBOROVÁ, *Inferring the origin of an epidemic with a dynamic message-passing algorithm*, Phys. Rev. E, 90 (2014), p 012801.

[144] G. LUGOSI ET A. S. PEREIRA, *Finding the seed of uniform attachment trees*, arXiv :1801.01816, (2018).

[145] D. LUSSEAU, K. SCHNEIDER, O. J. BOISSEAU, P. HAASE, E. SLOOTEN ET S. M. DAWSON, *The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations*, Behavioral Ecology and Sociobiology, 54 (2003), p. 396–405.

[146] A. MAGNER, A. GRAMA, J. SREEDHARAN ET W. SZPANKOWSKI, *Recovery of vertex orderings in dynamic graphs*, dans Proceedings of the 2017 IEEE International Symposium on Information Theory (ISIT), IEEE, 2017, p. 1563–1567.

[147] A. MAGNER, A. GRAMA, J. K. SREEDHARAN ET W. SZPANKOWSKI, *TIMES : temporal information maximally extracted from structure*, dans Proceedings of the 2018 World Wide Web Conference (WWW), 2018, p. 389–398.

[148] S. MAHANTESH, S. IYENGAR, M. VIJESH, S. R. NAYAK ET N. SHENOY, *Prediction of arrival of nodes in a scale free network*, dans Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012), IEEE

Computer Society, 2012, p. 517–521.

[149] L. MASSOULIÉ, *Community detection thresholds and the weak ramanujan property*, dans Proceedings of the 46th Annual ACM Symposium on Theory of Computing, ACM, New York, 2014, p. 694–703.

[150] F. A. MASSUCCI, J. WHEELER, R. BELTRÁN-DEBÓN, J. JOVEN, M. SALES-PARDO ET R. GUIMERÀ, *Inferring propagation paths for sparsely observed perturbations on complex networks*, Sci. Adv., 2 (2017), p e1501638.

[151] N. MASUDA, M. A. PORTER ET R. LAMBIOTTE, *Random walks and diffusion on networks*, Phys. Rep., 716 (2017).

[152] S. MELNIK, A. HACKETT, M. A. PORTER, P. J. MUCHA ET J. P. GLEESON, *The unreasonable effectiveness of tree-based theory for networks with clustering*, Phys. Rev. E, 83 (2011), p 036112.

[153] M. MEZARD ET A. MONTANARI, *Information, Physics, and Computation*, Oxford University Press, Oxford, 2009.

[154] I. MIKLÓS, P. L. ERDŐS ET L. SOUKUP, *Towards random uniform sampling of bipartite graphs with given degree sequence*, Electron. J. Combin., 20 (2013), p P16.

[155] M. MITCHELL, *Complexity : A Guided Tour*, Oxford University Press, Oxford, 2009.

[156] M. MOLLOY ET B. A. REED, *A critical point for random graphs with a given degree sequence*, Rand. Struct. Alg., 6 (1995), p. 161–180.

[157] ——, *The size of the giant component of a random graph with a given degree sequence*, Comb. Probab. Comp., 7 (1998), p. 295–305.

[158] C. MOORE, *The computer science and physics of community detection : landscapes, phase transitions, and hardness*, arXiv :1702.00467, (2017).

[159] E. MOSSEL, J. NEEMAN ET A. SLY, *A proof of the block model threshold conjecture*, arXiv :1311.4115, (2013).

[160] ——, *Reconstruction and estimation in the planted partition model*, Probab. Theory Related Fields, 162 (2015), p. 431–461.

[161] C. MURPHY, A. ALLARD, E. LAURENCE, G. ST-ONGE ET L. J. DUBÉ, *Geometric evolution of complex networks with degree correlations*, Phys. Rev. E, 97 (2018), p 032309.

[162] R. R. NADAKUDITI ET M. E. J. NEWMAN, *Graph spectra and the detectability of community structure in networks*, Phys. Rev. Lett., 108 (2012), p 188701.

[163] R. R. NADAKUDITI ET M. E. J. NEWMAN, *Spectra of random graphs with arbitrary expected degrees*, Phys. Rev. E, 87 (2013), p 012803.

[164] S. NAVLAKHA ET C. KINGSFORD, *Network archaeology : uncovering ancient networks from present-day interactions*, PLoS Comput. Biol., 7 (2011), p e1001119.

[165] A. NEMATZADEH, E. FERRARA, A. FLAMMINI ET Y.-Y. AHN, *Optimal network modula-*

rity for information diffusion, Phys. Rev. Lett., 113 (2014), p 088701.

[166] M. E. J. NEWMAN, *Mixing patterns in networks*, Phys. Rev. E, 67 (2003), p 026126.

[167] ——, *Properties of highly clustered networks*, Phys. Rev. E, 68 (2003), p 026121.

[168] ——, *The structure and function of complex networks*, SIAM Rev., 45 (2003), p. 167–256.

[169] ——, *Modularity and community structure in networks*, Proc. Natl. Acad. Sci. U.S.A., 103 (2006), p. 8577–8582.

[170] ——, *Networks : An Introduction*, Oxford University Press, Oxford, 2010.

[171] ——, *Communities, modules and large-scale structure in networks*, Nat. Phys., 8 (2012), p 25.

[172] M. E. J. NEWMAN, *Community detection and graph partitioning*, Europhys. Lett., 103 (2013), p 28003.

[173] M. E. J. NEWMAN, *Spectral methods for community detection and graph partitioning*, Phys. Rev. E, 88 (2013), p 042822.

[174] ——, *Equivalence between modularity optimization and maximum likelihood methods for community detection*, Phys. Rev. E, 94 (2016), p 052315.

[175] M. E. J. NEWMAN ET A. CLAUSET, *Structure and inference in annotated networks*, Nat. Comm., 7 (2016), p 11863.

[176] M. E. J. NEWMAN ET M. GIRVAN, *Finding and evaluating community structure in networks*, Phys. Rev. E, 69 (2004), p 026113.

[177] M. E. J. NEWMAN ET G. REINERT, *Estimating the number of communities in a network*, Phys. Rev. Lett., 117 (2016), p 078301.

[178] M. E. J. NEWMAN, S. H. STROGATZ ET D. J. WATTS, *Random graphs with arbitrary degree distributions and their applications*, Phys. Rev. E, 64 (2001), p 026118.

[179] K. NOWICKI ET T. A. B. SNIJDERS, *Estimation and prediction for stochastic blockstructures*, J. Am. Stat. Assoc., 96 (2001), p. 1077–1087.

[180] S. C. OLHEDE ET P. J. WOLFE, *Network histograms and universality of blockmodel approximation*, Proc. Natl. Acad. Sci. U.S.A., 111 (2014), p. 14722–14727.

[181] R. OLIVEIRA ET J. SPENCER, *Connectivity transitions in networks with super-linear preferential attachment*, Internet Math., 2 (2005), p. 121–163.

[182] C. ORSINI, M. M. DANKULOV, A. JAMAKOVIC, P. MAHADEVAN, P. COLOMER-DE SIMÓN, A. VAHDAT, K. E. BASSLER, Z. TOROCZKAI, M. BOGUÑÁ, G. CALDARELLI *et al.*, *Quantifying randomness in real networks*, Nat. Commun., 6 (2015), p 8627.

[183] N. OTTER, M. A. PORTER, U. TILLMANN, P. GRINDROD ET H. A. HARRINGTON, *A roadmap for the computation of persistent homology*, EPJ Data Sci., 6 (2017), p 17.

[184] A. PARANJAPE, A. R. BENSON ET J. LESKOVEC, *Motifs in temporal networks*, dans Proceedings of the 10th ACM International Conference on Web Search and Data Mining,

ACM, 2017, p. 601–610.

[185] R. Pastor-Satorras, C. Castellano, P. Van Mieghem et A. Vespignani, *Epidemic processes in complex networks*, Rev. Mod. Phys., 87 (2015), p 925.

[186] A. Patania, F. Vaccarino et G. Petri, *Topological analysis of data*, EPJ Data Sci., 6 (2017), p 7.

[187] R. Patro et C. Kingsford, *Predicting protein interactions via parsimonious network history inference*, Bioinformatics, 29 (2013), p. i237–i246.

[188] R. Patro, E. Sefer, J. Malin, G. Marçais, S. Navlakha et C. Kingsford, *Parsimonious reconstruction of network evolution*, Algorithm Mol. Biol., 7 (2012), p 25.

[189] L. Peel, D. B. Larremore et A. Clauset, *The ground truth about metadata and community detection in networks*, Sci. Adv., 3 (2017), p e1602548.

[190] T. P. Peixoto, *Entropy of stochastic blockmodel ensembles*, Phys. Rev. E, 85 (2012), p 056122.

[191] ——, *Eigenvalue spectra of modular networks*, Phys. Rev. Lett., 111 (2013), p 098701.

[192] ——, *Parsimonious module inference in large networks*, Phys. Rev. Lett., 110 (2013), p 148701.

[193] ——, *Efficient Monte Carlo and greedy heuristic for the inference of stochastic block models*, Phys. Rev. E, 89 (2014), p 012804.

[194] ——, *The graph-tool python library*, 2014.

[195] ——, *Hierarchical block structures and high-resolution model selection in large networks*, Phys. Rev. X, 4 (2014), p 011047.

[196] ——, *Model selection and hypothesis testing for large-scale network models with overlapping groups*, Phys. Rev. X, 5 (2015), p 011033.

[197] ——, *Bayesian stochastic blockmodeling*, arXiv :1705.10225, (2017).

[198] G. Petri, P. Expert, F. Turkheimer, R. Carhart-Harris, D. Nutt, P. Hellyer et F. Vaccarino, *Homological scaffolds of brain functional networks*, J. R. Soc. Interface, 11 (2014), p 20140873.

[199] G. Petri, M. Scolamiero, I. Donato et F. Vaccarino, *Topological strata of weighted complex networks*, PLoS One, 8 (2013), p e66506.

[200] T. Pham, P. Sheridan et H. Shimodaira, *PAFit : a statistical method for measuring preferential attachment in temporal complex networks*, PLoS One, 10 (2015), p e0137796.

[201] ——, *Joint estimation of preferential attachment and node fitness in growing complex networks*, Sci. Rep., 6 (2016).

[202] J. W. Pinney, G. D. Amoutzias, M. Rattray et D. L. Robertson, *Reconstruction of ancestral protein interaction networks for the bZIP transcription factors*, Proc. Natl. Acad. Sci. U.S.A., 104 (2007), p. 20449–20453.

[203] M. A. PORTER, J.-P. ONNELA ET P. J. MUCHA, *Communities in networks*, Notices of the AMS, 56 (2009), p. 1082–1097.

[204] W. H. PRESS, *Numerical Recipes : The Art of Scientific Computing*, Cambridge University press, 3eme éd., 2007.

[205] M. Z. RÁCZ, S. BUBECK *et al.*, *Basic models and questions in statistical network analysis*, Stat. Surv., 11 (2017), p. 1–47.

[206] U. N. RAGHAVAN, R. ALBERT ET S. KUMARA, *Near linear time algorithm to detect community structures in large-scale networks*, Phys. Rev. E, 76 (2007), p 036106.

[207] J. REICHARDT ET S. BORNHOLDT, *Statistical mechanics of community detection*, Phys. Rev. E, 74 (2006), p 016110.

[208] J. REICHARDT ET M. LEONE, *(Un)detectable cluster structure in sparse networks*, Phys. Rev. Lett., 101 (2008), p 078701.

[209] C. ROBERT, *The Bayesian Choice : From Decision-Theoretic Foundations to Computational Implementation*, Springer, New York, 2007.

[210] M. P. ROMBACH, M. A. PORTER, J. H. FOWLER ET P. J. MUCHA, *Core-periphery structure in networks*, SIAM J. Appl. Math., 74 (2014), p. 167–190.

[211] P. RONHOVDE ET Z. NUSSINOV, *Local resolution-limit-free potts model for community detection*, Phys. Rev. E, 81 (2010), p 046114.

[212] M. ROSVALL ET C. T. BERGSTROM, *An information-theoretic framework for resolving community structure in complex networks*, Proc. Natl. Acad. Sci. U.S.A., 104 (2007), p. 7327–7331.

[213] M. ROSVALL ET C. T. BERGSTROM, *Maps of random walks on complex networks reveal community structure*, Proc. Natl. Acad. Sci. U.S.A., 105 (2008), p. 1118–1123.

[214] A. ROXANA PAMFIL, S. D. HOWISON, R. LAMBIOTTE ET M. A. PORTER, *Relating modularity maximization and stochastic block models in multilayer networks*, arXiv :1804.01964, (2018).

[215] S. V. SCARPINO, A. ALLARD ET L. HÉBERT-DUFRESNE, *The effect of a prudent adaptive behaviour on disease transmission*, Nat. Phys., 12 (2016), p 1042.

[216] M. T. SCHAUB, J.-C. DELVENNE, M. ROSVALL ET R. LAMBIOTTE, *The many facets of community detection in complex networks*, Appl. Netw. Sci., 2 (2017), p 4.

[217] S. B. SEIDMAN, *Network structure and minimum degree*, Soc. Networks, 5 (1983), p. 269–287.

[218] C. SESHADHRI, T. G. KOLDA ET A. PINAR, *Community structure and scale-free collections of Erdős-Rényi graphs*, Phys. Rev. E., 85 (2012), p 056109.

[219] D. SHAH ET T. ZAMAN, *Rumors in a network : who's the culprit ?*, IEEE T. Inform. Theory, 57 (2011), p. 5163–5181.

[220] S. Shai, N. Stanley, C. Granell, D. Taylor et P. J. Mucha, *Case studies in network community detection*, arXiv :1705.02305, (2017).

[221] C. E. Shannon, *A mathematical theory of communication*, Bell Syst. Tech. J., 27 (1948), p. 379–423.

[222] H. A. Simon, *Models of Man ; Social and Rational*, Wiley, New York, 1957.

[223] A. Sizemore, C. Giusti et D. S. Bassett, *Classification of weighted networks through mesoscale homological features*, J. Complex Netw., 5 (2017), p. 245–273.

[224] T. A. Snijders et K. Nowicki, *Estimation and prediction for stochastic blockmodels for graphs with latent block structure*, Journal of Classification, 14 (1997), p. 75–100.

[225] G. St-Onge, J.-G. Young, E. Laurence, C. Murphy et L. J. Dubé, *Susceptible-infected-susceptible dynamics on the rewired configuration model*, arXiv :1701.01740, (2017).

[226] ——, *Phase transition of the susceptible-infected-susceptible dynamics on time-varying configuration model networks*, Phys. Rev. E, 97 (2018), p 022305.

[227] D. L. Stein et C. M. Newman, *Spin Glasses and Complexity*, Princeton University Press, Princeton, NJ, Princeton, 2013.

[228] B. Stolz, H. Harrington et M. A. Porter, *The topological shape of Brexit*, arXiv :1610.00752, (2016).

[229] P. Šulc et L. Zdeborová, *Belief propagation for graph partitioning*, J. Phys. A, 43 (2010), p 285003.

[230] S. Thurner, R. Hanel, B. Liu et B. Corominas-Murtra, *Understanding Zipf's law of word frequencies through sample-space collapse in sentence formation*, J. R. Soc. Interface, 12 (2016), p 20150330.

[231] G. Tibély et J. Kertész, *On the equivalence of the label propagation method of community detection and a potts model approach*, Physica A, 387 (2008), p. 4982–4984.

[232] V. A. Traag, P. Van Dooren et Y. Nesterov, *Narrow scope for resolution-limit-free community detection*, Phys. Rev. E, 84 (2011), p 016114.

[233] S. van der Pas et A. van der Vaart, *Bayesian community detection*, arXiv :1608.04242, (2016).

[234] N. Veldt, D. Gleich et A. Wirth, *A correlation clustering framework for community detection*, (2018), p. 439–448.

[235] G. Ver Steeg, C. Moore, A. Galstyan et A. Allahverdyan, *Phase transitions in community detection : A solvable toy model*, Europhys. Lett., 106 (2014), p 48004.

[236] S. Wasserman et C. Anderson, *Stochastic a posteriori blockmodels : Construction and assessment*, Soc. Networks, 9 (1987), p. 1–36.

[237] W. Weaver, *Science and complexity*, American Scientist, 36 (1948), p 536.

[238] H. C. White, S. A. Boorman et R. L. Breiger, *Social structure from multiple networks.*

Blockmodels of roles and positions, Am. J. Sociol., (1976), p. 730–780.

[239] S. WHITE ET P. SMYTH, *A spectral clustering approach to finding communities in graphs*, dans Proceedings of the 2005 SIAM International Conference on Data Mining, SIAM, Philadelphia, PA, 2005, p. 274–285.

[240] WILF, H.S., *generatingfunctionology*, Academic Press inc., 1994.

[241] C. WIUF, M. BRAMEIER, O. HAGBERG ET M. P. STUMPF, *A likelihood approach to analysis of network data*, Proc. Natl. Acad. Sci. U.S.A., 103 (2006), p. 7566–7570.

[242] Z. WU, G. MENICHETTI, C. RAHMEDE ET G. BIANCONI, *Emergent complex network geometry*, Sci. Rep., 5 (2015), p 10073.

[243] K. XIA ET G.-W. WEI, *Persistent homology analysis of protein structure, flexibility, and folding*, Int. J. Numer. Methods Biomed. Eng., 30 (2014), p. 814–844.

[244] J. YANG ET J. LESKOVEC, *Overlapping communities explain core-periphery organization of networks*, Proc. IEEE, 102 (2014), p. 1892–1902.

[245] J.-G. YOUNG, *De la détection de la structure communautaire des réseaux complexes*, Mémoire de maîtrise, Université Laval, 2014.

[246] J.-G. YOUNG, A. ALLARD, L. HÉBERT-DUFRESNE ET L. J. DUBÉ, *A shadowing problem in the detection of overlapping communities : Lifting the resolution limit through a cascading procedure*, PLoS One, 10 (2015), p e0140133.

[247] J.-G. YOUNG, P. DESROSIERS, L. HÉBERT-DUFRESNE, E. LAURENCE ET L. J. DUBÉ, *Finite-size analysis of the detectability limit of the stochastic block model*, Phys. Rev. E, 95 (2017), p 062304.

[248] J.-G. YOUNG, L. HÉBERT-DUFRESNE, A. ALLARD ET L. J. DUBÉ, *Growing networks of overlapping communities with internal structure*, Phys. Rev. E, 94 (2016), p 022317.

[249] J.-G. YOUNG, L. HÉBERT-DUFRESNE, E. LAURENCE, C. MURPHY, G. ST-ONGE ET P. DESROSIERS, *Network archaeology : phase transition in the recoverability of network history*, arXiv :1803.09191, (2018).

[250] J.-G. YOUNG, G. PETRI, F. VACCARINO ET A. PATANIA, *Construction of and efficient sampling from the simplicial configuration model*, Phys. Rev. E, 96 (2017), p 032312.

[251] J.-G. YOUNG, G. ST-ONGE, P. DESROSIERS ET L. J. DUBÉ, *Universality of the stochastic block model*, arXiv :1806.04214, (2018).

[252] L. ZDEBOROVÁ ET F. KRZAKALA, *Statistical physics of inference : Thresholds and algorithms*, Adv. Phys., 65 (2016), p. 453–552.

[253] P. ZHANG, *Evaluating accuracy of community detection using the relative normalized mutual information*, J. Stat. Mech., (2015), p P11006.

[254] P. ZHANG ET C. MOORE, *Scalable detection of statistically significant communities and hierarchies, using message passing for modularity*, Proc. Natl. Acad. Sci. U.S.A., 111 (2014),

p. 18144–18149.

[255] P. Zhang, C. Moore et M. E. J. Newman, *Community detection in networks with unequal groups*, Phys. Rev. E, 93 (2016), p 012303.

[256] X. Zhang et B. M. E. Moret, *Refining transcriptional regulatory networks using network evolutionary models and gene histories*, Algorithm Mol. Biol., 5 (2010), p 1.

[257] X. Zhang, R. R. Nadakuditi et M. E. J. Newman, *Spectra of random graphs with community structure and arbitrary degrees*, Phys. Rev. E, 89 (2014), p 042816.

[258] G.-M. Zhu, H. Yang, R. Yang, J. Ren, B. Li et Y.-C. Lai, *Uncovering evolutionary ages of nodes in complex networks*, Eur. Phys. J. B, 85 (2012), p 106.

[259] K. Zuev, O. Eisenberg et D. Krioukov, *Exponential random simplicial complexes*, J. Phys. A, 48 (2015), p 465002.