Percolation sur graphes aléatoires
– modélisation et description analytique –

Thèse

Antoine Allard

Doctorat en physique
Philosophiæ doctor (Ph.D.)

Québec, Canada

© Antoine Allard, 2014

Résumé

Les graphes sont des objets mathématiques abstraits utilisés pour modéliser les interactions entre les éléments constitutifs des systèmes complexes. Cette utilisation est motivée par le fait qu'il existe un lien fondamental entre la structure de ces interactions et les propriétés macroscopiques de ces systèmes. La théorie de la percolation offre un paradigme de choix pour analyser la structure de ces graphes, et ainsi mieux comprendre les conditions dans lesquelles ces propriétés émergent.

Les interactions dans une grande variété de systèmes complexes partagent plusieurs propriétés structurelles *universelles*, et leur incorporation dans un cadre théorique unique demeure l'un des principaux défis de l'étude des systèmes complexes. Exploitant une approche *multitype*, une idée toute simple mais étonnamment puissante, nous avons unifié l'ensemble des modèles de percolation sur graphes aléatoires connus en un même cadre théorique, ce qui en fait le plus général et le plus réaliste proposé à ce jour. Bien plus qu'une simple compilation, le formalisme que nous proposons augmente significativement la complexité des structures pouvant être reproduites et, de ce fait, ouvre la voie à plusieurs nouvelles avenues de recherche.

Nous illustrons cette assertion notamment en utilisant notre modèle pour valider et formaliser certaines intuitions inspirées de résultats empiriques. Dans un premier temps, nous étudions comment la structure en réseau de certains systèmes complexes (ex. réseau de distribution électrique, réseau social) facilite leur surveillance, et par conséquent leur éventuel contrôle. Dans un second temps, nous explorons la possibilité d'utiliser la décomposition en couches "k-core" en tant que structure effective des graphes extraits des systèmes complexes réels. Enfin, nous utilisons notre modèle pour identifier les conditions pour lesquelles une nouvelle stratégie d'immunisation contre des maladies infectieuses est la stratégie optimale.

Abstract

Graphs are abstract mathematical objects used to model the interactions between the elements of complex systems. Their use is motivated by the fact that there exists a fundamental relationship between the structure of these interactions and the macroscopic properties of these systems. The structure of these graphs is analyzed within the paradigm of percolation theory whose tools and concepts contribute to a better understanding of the conditions for which these emergent properties appear.

The underlying interactions of a wide variety of complex systems share many *universal* structural properties, and including these properties in a unified theoretical framework is one of the main challenges of the science of complex systems. Capitalizing on a multitype approach, a simple yet powerful idea, we have unified the models of percolation on random graphs published to this day in a single framework, hence yielding the most general and realistic framework to date. More than a mere compilation, this framework significantly increases the structural complexity of the graphs that can now be mathematically handled, and, as such, opens the way to many new research opportunities.

We illustrate this assertion by using our framework to validate hypotheses hinted at by empirical results. First, we investigate how the network structure of some complex systems (e.g., power grids, social networks) enhances our ability to monitor them, and ultimately to control them. Second, we test the hypothesis that the "k-core" decomposition can act as an effective structure of graphs extracted from real complex systems. Third, we use our framework to identify the conditions for which a new immunization strategy against infectious diseases is optimal.

Table des matières

Liste des tableaux

Liste des figures

À qui rêve et ose.

And that the knowledge is just to put into correct framework the wonder that nature is.

Richard P. Feynman, Italie, 1964

Remerciements

Bien qu'une thèse soit en principe une œuvre individuelle, elle est en pratique le fruit d'un effort collectif. De près ou de loin, plusieurs personnes ont contribué à l'aboutissement de ce doctorat, une entreprise audacieuse qui germa d'un songe lors d'un automne cégepien. Je prends donc quelques lignes pour humblement les remercier.

Il est d'usage de débuter cette section avec le directeur de recherche. Dans ce cas-ci, cet ordre n'est cependant pas dicté par la coutume, mais bien par le mérite. Ainsi, je remercie Louis J. Dubé pour sa confiance, sa curiosité, sa patience et son dévouement. Ayant également été mon directeur de recherche à la maîtrise, je lui dois une grande partie de ce que je suis aujourd'hui comme jeune chercheur. Cette thèse porte sa couleur, et j'espère avoir réussi à l'élever au standard de qualité qu'il exige du travail de ses étudiants et collaborateurs. En plus de laisser à ses étudiants la latitude pour explorer la moindre idée, il sait réunir au sein d'une même équipe des individus dont la contribution permet au groupe d'être bien plus que la somme de chacun des membres. Si un jour j'ai la chance de participer à la formation de jeunes maîtres, mon approche et mon attitude seront fortement inspirées des siennes.

J'aimerais remercier messieurs Yvan Saint-Aubin, Pierre Mathieu et Patrick Desrosiers pour avoir accepté de siéger sur le comité d'évaluation de cette thèse. Le sujet qui y est traité sort quelque peu des sentiers battus et, conséquemment, possède peu de points communs avec la physique théorique « classique », qui auraient offert au lecteur quelques points de repère. J'ai mis beaucoup d'efforts pour que son contenu soit présenté clairement et pour que les nouveaux concepts soient bien expliqués. J'espère que cette lecture vous sera enrichissante et agréable.

J'aimerais lever mon chapeau aux membres de la confrérie. Je salue d'abord les gars de Dynamica::réseaux : Pierre-André pour sa rigueur et sa gentillesse ; Laurent pour son insatiable appétit et son intarissable imagination ; Jean-Gabriel pour sa curiosité et son émerveillement ; Vincent pour sa constance et sa précision. Mes travaux n'auraient jamais eu leur envergure si ce n'étaient de vos conseils et de vos idées. Merci d'avoir supporté mes humeurs, et j'espère vous avoir été utile à l'occasion. Je salue ensuite les gars de lumière. Même si nos travaux ne se recoupaient que très peu, ce fut un plaisir de vous côtoyer quotidiennement, et je reste fasciné par la quantité de trucs parfois utiles, parfois absolument inutiles que vous m'avez fait découvrir. Ce fut un plaisir de travailler avec vous tous, et si nos chemins doivent se séparer, je suivrai vos exploits à distance avec un plaisir toujours renouvelé.

J'aimerais également remercier mes parents pour le reportage sur Voyager et l'étincelle dans leurs yeux qui me donne un sentiment de pertinence ; Olivier, Jean-Philippe et Vincent pour la musique, bien sûr, et le rappel incessant de ma propre insignifiance ; Marc Olivier pour l'inspiration première et la rigueur de pensée ; Jérôme pour le recul et la tête hors de l'eau ; Pierre pour le jeu et le mépris partagé de l'idiotie ; Sarah pour l'intensité et l'humour ; et Catherine pour l'ardeur et le coup de pied initial. Enfin, je remercie tous ceux chez qui l'herbe semble plus verte, et qui par leur passion et leur talent m'incitent à aller au bout de moi-même.

Et un merci tout spécial à Gabrielle qui me montre ce que j'ai tant tardé à voir.

Aide-mémoire

Afin de faciliter la lecture de ce document, nous avons colligé dans le tableau 1 la plupart des termes techniques, en français et en anglais, couramment utilisés pour que le lecteur puisse s'y référer au besoin.

Terme français	Terme anglais	Description
nœud	node	Unité fondamentale d'un graphe correspondant à un des éléments en interaction dans un système complexe. Aussi parfois appelé *sommet*.
lien	edge (link)	Unité fondamentale d'un graphe correspondant à une interaction entre deux éléments d'un système complexe. Aussi parfois appelé *arête*.
lien bidirectionnel	undirected edge	Lien pouvant être parcouru dans l'un ou l'autre de ses sens.
lien unidirectionnel	directed edge	Lien ne pouvant être parcouru que dans un de ses sens.
motif	hyperedge (motif)	Généralisation du concept de lien. Un motif lie plus de deux nœuds selon une configuration de liens donnée.
degré	degree	Nombre de liens attachés à un nœud. On distingue les degrés *entrant* et *quittant* (*incoming degree* et *outgoing degree*) en présence de liens unidirectionnels.
degré sortant	excess degree	Nombre de liens par lesquels il est possible de quitter un nœud qui a été atteint via un de ses liens (c.-à-d. égal au degré du nœud moins 1).
voisinage	neighborhood	Ensemble des nœuds avec lesquels un nœud donné partage un lien ou un motif (aussi appelé *premier* voisinage). Le *second* voisinage correspond à l'union des premiers voisinages des nœuds appartenant au premier voisinage d'un nœud donné.
composante	component	Groupe de nœuds isolés du reste du graphe où chacun des membres est le voisin (premier, second, etc.) des autres nœuds du groupe. Une composante est dite *géante* si le nombre de nœuds qu'elle contient est une quantité extensive. Autrement, il s'agit d'une *petite* composante.
matrice d'adjacence	adjacency matrix	Représentation matricielle de la structure d'un graphe dans laquelle l'élément A_{ij} indique le nombre de liens quittant le i-ième nœud vers le j-ième nœud.
distribution des degrés	degree distribution	Distribution du degré des nœuds dans un graphe.
agrégation	clustering	Tendance qu'ont les nœuds d'un graphe à s'agglomérer en groupes où les connexions sont redondantes (ex. le premier voisin d'un nœud peut également être son second voisin par l'entremise d'un troisième nœud). Au premier ordre, on mesure l'agrégation par la présence de triangles.

TABLE 1 – Glossaire des termes techniques couramment utilisés.

Contributions scientifiques

Publications

Percolation sur graphes aléatoires

- *A general and exact approach to percolation on random graphs*
  A. Allard, L. Hébert-Dufresne, J.-G. Young, and L. J. Dubé
  en préparation

- *Coexistence of phases and the observability of random graphs*
  A. Allard, L. Hébert-Dufresne, J.-G. Young, and L. J. Dubé
  Phys. Rev. E, **89** 022801 (2014) [9]

- *Percolation on random networks with arbitrary k-core structure*
  L. Hébert-Dufresne [1], **A. Allard**[1], J.-G. Young, and L. J. Dubé
  Phys. Rev. E 88, 062820 (2013) [92]

- *Bond percolation on a class of correlated and clustered random graphs*
  A. Allard, L. Hébert-Dufresne, P.-A. Noël, V. Marceau, and L. J. Dubé
  J. Phys. A 45, 405005 (2012) [6]

- *Exact solution of bond percolation on small arbitrary graphs*
  A. Allard, L. Hébert-Dufresne, P.-A. Noël, V. Marceau, and L. J. Dubé
  EPL 98, 16001 (2012) [7]

Détection de communautés

- *A system-level model for the microbial regulatory genome*
  A. N. Brooks, D. J. Reiss, **A. Allard**, W.-J. Wu, D. M. Salvanha, C. L. Plaisier, S. Chandrasekaran, M. Pan, A. Kaur, and N. S. Baliga
  soumis à Molecular Systems Biology

[1] Ces auteurs ont contribué également à cet article.

- *Unveiling hidden communities through cascading detection on network structures*
  J.-G. Young, **A. Allard**, L. Hébert-Dufresne, and L. J. Dubé
  en préparation

Propriétés universelles des systèmes complexes réels

- *Complex networks are an emerging property of hierarchical preferential attachment*
  L. Hébert-Dufresne, E. Laurence, **A. Allard**, J.-G. Young, and L. J. Dubé
  en préparation
- *Universal growth constraints of human systems*
  L. Hébert-Dufresne, **A. Allard**, J.-G. Young, and L. J. Dubé
  en préparation
- *Structural preferential attachment : Stochastic process for the growth of scale-free, modular and self-similar systems*
  L. Hébert-Dufresne, **A. Allard**, V. Marceau, P.-A. Noël, and L. J. Dubé
  Phys. Rev. E 85, 026108 (2012) [89]
- *Structural preferential attachment : Network organization beyond the link*
  L. Hébert-Dufresne, **A. Allard**, V. Marceau, P.-A. Noël, and L. J. Dubé
  Phys. Rev. Lett. 107, 158702 (2011) [88]

Processus stochastiques sur graphes

- *The Social Zombie : Modelling undead outbreaks on social networks*
  L. Hébert-Dufresne, P.-A. Noël, V. Marceau, **A. Allard**, and L. J. Dubé
  R. Smith? (Ed.), Les Presses de l'Université d'Ottawa, *à paraître en octobre 2014*
- *Epidemics on contact networks : a general stochastic approach*
  P.-A. Noël, **A. Allard**, L. Hébert-Dufresne, V. Marceau, and L. J. Dubé
  J. Math. Biol. (2013) [150]
- *Global efficiency of local immunization of complex networks*
  L. Hébert-Dufresne[1], **A. Allard**[1], J.-G. Young[1], and L. J. Dubé
  Sci. Rep. 3, 2171 (2013) [91]
- *Propagation on networks : an exact alternative perspective*
  P.-A. Noël, **A. Allard**, L. Hébert-Dufresne, V. Marceau, and L. J. Dubé
  Phys. Rev. E 85, 031118 (2012) [149]
- *Modeling the dynamical interaction between epidemics on overlay networks*
  V. Marceau, P.-A. Noël, L. Hébert-Dufresne, **A. Allard**, and L. J. Dubé
  Phys. Rev. E 84, 026105 (2011) [117]

- *Adaptive networks : Coevolution of disease and topology*
  V. Marceau, P.-A. Noël, L. Hébert-Dufresne, **A. Allard**, and L. J. Dubé
  Phys. Rev. E 82, 036116 (2010) [116]

- *Propagation dynamics on networks featuring complex topologies*
  L. Hébert-Dufresne, P.-A. Noël, V. Marceau, **A. Allard**, and L. J. Dubé
  Phys. Rev. E 82, 036115 (2010) [93]

Présentations

(sélection, le nom du conférencier est souligné)

- *Percolation on clustered and correlated random graphs : General formalism and applications* (affiche)
  A. Allard, L. Hébert-Dufresne, J.-G. Young, and L. J. Dubé
  International School and Conference on Network Science, Copenhague, Danemark, 2013

- *Hard-core random networks as an effective model of bond percolation on real networks* (orale)
  L. Hébert-Dufresne, **A. Allard**, J.-G. Young, and L. J. Dubé
  International School and Conference on Network Science, Copenhague, Danemark, 2013

- *Bond and site percolation on clustered and correlated random graphs* (orale)
  A. Allard, L. Hébert-Dufresne, J.-G. Young, and L. J. Dubé
  Joint CRM-Imperial College School and Workshop in Complex Systems, Barcelone, Espagne, 2013

- *Unveiling hidden communities through cascading detection on network structures* (orale)
  J.-G. Young, **A. Allard**, L. Hébert-Dufresne, and L. J. Dubé
  2$^{\text{nd}}$ International Conference on Complex Sciences, Santa Fe, Nouveau-Mexique, 2012

- *Exact solution of bond percolation on small arbitrary graphs* (orale)
  A. Allard, L. Hébert-Dufresne, P.-A. Noël, V. Marceau, and L. J. Dubé
  International School and Conference on Network Science, Evanston, Illinois, 2012

- *Using network organization to hinder propagation in structured populations* (affiche)
  L. Hébert-Dufresne, **A. Allard**, J.-G. Young, and L. J. Dubé
  International School and Conference on Network Science, Evanston, Illinois, 2012

- *Multitype modular networks as a model of clustered social networks* (affiche)
  A. Allard, P.-A. Noël, L. Hébert-Dufresne, V. Marceau, and L. J. Dubé
  International School and Conference on Network Science, Boston & Cambridge, Massachusetts, 2010

Prologue

Plusieurs des grands succès de la science moderne sont les fruits de l'approche *réductionniste* dans laquelle l'apparente complexité d'un système physique ou biologique peut être levée en en étudiant les parties. Par exemple, le fonctionnement du corps humain nous est principalement révélé par notre compréhension de ses différents organes et, similairement, notre compréhension de ceux-ci résulte des connaissances acquises au sujet de leurs propres éléments constitutifs (ex. cellules). Un autre exemple nous vient de la physique atomique et de son évolution, de la fin du XIX^{e} siècle à nos jours. La mécanique quantique, l'énergie nucléaire, le transistor et, récemment, la « sacralisation » du modèle standard suivant la confirmation expérimentale de l'existence du boson de Higgs témoignent des succès obtenus à mesure que nous creusons vers l'infiniment petit.

Il existe cependant une limite à la perspective que nous offre le réductionnisme pour déchiffrer et comprendre le monde qui nous entoure. Pensons simplement au cerveau humain dont les facultés vont bien au-delà de la somme de celles des milliards de neurones qui le constituent. De fait, lorsque considérés à leur échelle globale, plusieurs systèmes dits *complexes* arborent des propriétés *émergentes* absentes chez leurs éléments constitutifs et qui résultent plutôt des interactions entre ceux-ci. Pensons notamment à la *résilience* des milieux écologiques face aux changements de leur environnement (ex. disparition d'espèces), à la *stabilité* des réseaux de distribution électrique, à la *navigabilité* d'Internet sans une connaissance explicite de sa structure globale, ou à la *susceptibilité* d'une population face à une nouvelle souche d'un agent pathogène (ex. Influenza).

Bien qu'ils soient de nature et d'échelle diverses, ces systèmes complexes partagent une caractéristique fondamentale : leurs *fonctions*, ou leurs propriétés *macroscopiques*, sont directement liées à la *structure* des interactions entre leurs parties constituantes. Au besoin d'étudier les éléments individuels s'ajoute donc la nécessité d'élucider la nature et l'organisation de leurs interactions. Or, le développement fulgurant de l'informatique et des techniques de mesure expérimentales a récemment ouvert la voie à l'étude quantitative d'un grand nombre de systèmes complexes, pour lesquels seules des études principalement qualitatives étaient possibles. Ces recherches ont révélé qu'en plus d'avoir en commun ce lien structure-fonctions, les interactions sous-jacentes à plusieurs de ces systèmes partagent des propriétés structurelles *universelles*. Ceci suggère que, malgré leur apparente dissimilitude, ces systèmes peuvent être étudiés dans un même cadre conceptuel.

La physique n'est pas étrangère à l'étude de systèmes composés d'un grand nombre d'éléments en interaction, un exemple probant étant le pont qu'elle a établi entre les lois de la thermodynamique et la nature atomique de la matière. Ainsi, s'inspirant de concepts et de techniques de la *physique statistique* et de la *dynamique non linéaire*, il est possible de construire des modèles mathématiques de systèmes complexes pour tenter d'expliquer l'origine de leurs propriétés structurelles universelles, et mieux comprendre la relation structure-fonctions.

Mathématiquement, ces interactions sont représentées à l'aide de *graphes* : des structures abstraites formées de points (les *nœuds*) liés les uns aux autres (les *liens*). Due à Euler, cette représentation offre une perspective dépouillée du superflu et ramenée à l'essentiel : qui interagit avec qui ? Or, plusieurs fonctions, ou propriétés macroscopiques, de systèmes complexes sont liées à la *connectivité* des graphes sous-tendus par leurs interactions. Étudier la connectivité d'un graphe correspond notamment à déterminer les conditions pour lesquelles le graphe possède une composante extensive [2], ainsi que la taille de cette composante (c.-à-d. le nombre de nœuds qu'elle contient), lorsqu'elle existe. On peut également se demander si la connectivité d'un graphe est *robuste* à la défaillance de ses éléments : quelle fraction des nœuds ou des liens doit être retirée pour que la composante extensive disparaisse ? Les systèmes de communication, tels qu'Internet, ou de distribution, tels que les réseaux électriques, sont des exemples de systèmes complexes pour lesquels l'importance de la connectivité est manifeste. Plus subtile, la diffusion d'opinions est un phénomène dépendant également de la connectivité du graphe sous-tendu par les interactions sociales.

La connectivité s'étudie grâce, entre autres, au paradigme de la *percolation sur graphes* : une théorie conceptuellement simple, mais suffisamment riche pour en saisir toute la complexité. En fait, la percolation est la théorie la plus simple possédant une transition de phase : le passage soudain d'un état macroscopique à un autre (en l'occurrence, l'apparition d'une composante extensive). Plus précisément, une partie importante de la modélisation des graphes sous-jacents aux systèmes complexes se fait à l'aide de modèles de *graphes aléatoires*, dont la structure n'est pas décrite par un graphe particulier, mais bien par un ensemble dans lequel chaque graphe apparaît avec une certaine probabilité.

Bien que la structure des graphes extraits de systèmes complexes réels n'ait typiquement rien d'aléatoire, ces modèles possèdent plusieurs avantages justifiant leur utilisation. Premièrement, malgré qu'ils considèrent des graphes dont la structure est beaucoup plus simple que

[2] Une composante est un ensemble de nœuds liés de sorte que chaque nœud peut être rejoint à partir de n'importe quel autre nœud. Une composante est extensive si la fraction des nœuds du graphe qu'elle contient tend vers une fraction constante lorsque la taille du graphe tend vers l'infini. Ces concepts sont approfondis au chapitre 1.

celle des graphes extraits de systèmes réels, ils reproduisent qualitativement le comportement observé chez ces systèmes réels (c.-à-d. la transition de phase), et leurs prédictions sont parfois étonnamment justes. Deuxièmement, plusieurs des propriétés des modèles de graphes aléatoires peuvent être obtenues analytiquement, ce qui est un attribut rare. En effet, les systèmes complexes méritent bien leur épithète, ce qui a pour conséquence la nécessité de faire appel à des simulations numériques pour en faire l'étude. Or, bien que les études purement numériques aient leur utilité, elles n'offrent pas toujours un éclairage complet sur les mécanismes sous-jacents et les propriétés fondamentales des systèmes étudiés. Avoir à notre disposition une classe de modèles suffisamment réalistes pouvant être résolus analytiquement est donc une occasion à saisir. Troisièmement, les graphes aléatoires possèdent une structure dont l'entropie est maximale : cette structure est aléatoire à tous égards à l'exception des contraintes qui lui sont explicitement imposées. Ceci permet donc d'étudier dans un « environnement contrôlé » l'impact qu'ont différentes propriétés structurelles locales sur les propriétés macroscopiques des graphes telles que leur connectivité.

Dans cette thèse, nous présentons une classe très générale de graphes aléatoires possédant plusieurs propriétés structurelles universelles présentes dans les graphes extraits de systèmes complexes réels. Cette classe rassemble l'ensemble des modèles publiés à ce jour sous une même approche unifiée (c.-à-d. que ces modèles ne sont que des cas particuliers) et améliore considérablement la complexité des graphes pouvant être modélisés. Nous présentons également un formalisme mathématique permettant d'étudier la connectivité de ces graphes de façon exacte. La nature générale de notre approche lui confère le rôle de *laboratoire théorique*, un rôle que nous illustrons à l'aide de plusieurs exemples dans lesquels nous étudions l'impact qu'ont certaines propriétés structurelles locales sur la connectivité globale des graphes.

Le matériel original présenté dans cette thèse a déjà fait l'objet de plusieurs publications et c'est sous la forme de ces articles qu'il est exposé. Ainsi, les chapitres 2 à 5 et les annexes A et B contiennent essentiellement le texte original des articles, précédés d'un avant-propos offrant une brève présentation et expliquant comment chaque article s'inscrit dans le cadre de cette thèse. Un tableau synthèse résumant les différents objets mathématiques définis est également offert pour chaque chapitre. Les chapitres 2 à 4 présentent le cadre théorique que nous avons développé, et le chapitre 5, de même que les annexes A et B, sont des exemples d'application de notre approche générale. Puisque le ton de ces articles s'écarte quelque peu du style pédagogique propre à une thèse, nous offrons au chapitre 1 une revue en profondeur des concepts et outils mathématiques utilisés, afin que le lecteur en acquiert une compréhension solide. Nous énumérons et discutons des pistes intéressantes pour la poursuite de ces recherches dans la conclusion. Enfin, l'annexe C contient quelques remarques sur la convergence des simulations numériques, et l'annexe D décrit les méthodes numériques utilisées pour résoudre les équations du formalisme théorique.

Chapitre 1

Percolation, théorie des graphes et modélisation

Nous présentons dans ce chapitre les concepts et outils mathématiques sur lesquels sont fondés les travaux faisant l'objet des prochains chapitres. Ces chapitres étant essentiellement construits à partir du texte de publications scientifiques, leur ton s'écarte quelque peu du style pédagogique propre à une thèse. Dès lors, nous profitons de ce chapitre, d'une part, pour présenter les modèles fondateurs sur lesquels sont érigés les travaux doctoraux, et d'autre part, pour en offrir une analyse en profondeur afin que le lecteur acquiert une compréhension solide lorsque nous entrerons dans le vif du sujet.

1.1 Terminologie et concepts fondamentaux

Avant de s'attaquer aux modèles mathématiques, nous présentons sommairement la terminologie de même que quelques concepts fondamentaux qui nous serons utiles tout au long de cette thèse. Nous nous contenterons de définitions *opérationnelles* qui permettront une bonne compréhension intuitive des concepts. Le lecteur intéressé pourra se référer à tout bon manuel sur la théorie de la percolation et des phénomènes critiques (par exemple [42, 182, 185, 186]) de même que sur le théorie des graphes (par exemple [29, 55, 198]) pour des définitions plus formelles.

1.1.1 Percolation sur graphes aléatoires

Précisons d'abord ce que nous entendons par percolation sur graphes aléatoires. La plupart des ouvrages de référence introduisent la théorie de la percolation à l'aide du modèle suivant : soit une grille carrée de $N \to \infty$ cases dans laquelle chaque case (ou site) peut être colorée indépendamment avec une probabilité p. Pour de faibles valeurs de p, les cases colorés sont typiquement isolées et, à mesure que p augmente, des amas de cases colorées adjacentes se forment et croissent de façon continue et monotone. Or, à une valeur précise $p = p_c$, cette croissance cesse d'être monotone, et un amas composé d'une quantité extensive[1] de cases apparaît abruptement. Ce changement qualitatif dans l'organisation des amas correspond à une *transition de phase*, et ce modèle de cases colorées est l'un des plus simples à posséder une telle transition. Traditionnellement, la théorie de la percolation s'intéresse à l'organisation de ces amas en fonction de la probabilité p (ex. leur nombre, leur taille, leur forme). En particulier, autour du point critique p_c, la plupart des quantités décrivant l'organisation des amas arborent une *invariance d'échelle* et sont décrites par un ensemble d'exposants critiques universels. Par exemple, le paramètre d'ordre $\mathcal{S}$, correspondant à la fraction des cases appartenant à l'amas dont la taille est extensive, se comporte selon $\mathcal{S} \propto (p - p_c)^{\beta}$, où β est l'exposant critique, lorsque p est légèrement supérieur à p_c.

Ce modèle classique de la théorie de la percolation a été décliné en plusieurs versions au fil des années [66, 179][2]. Certaines versions changent la configuration de la grille (ex. grille triangulaire, de type Kagome [121]). D'autres considèrent des « hypergrilles » occupant des espaces à plus de deux dimensions (ex. grille « hypercubique » en 4 dimensions ou grille de Bethe [24]). Enfin, certaines versions s'intéressent plutôt à des grilles de points (aussi appelés nœuds) dans lesquelles ce sont les liens entre deux points adjacents qui existent indépendamment avec une probabilité p. On parle alors de *percolation par liens* par opposition à la *percolation par sites* présentée ci-dessus[3]. Alors qu'il est manifeste que l'une ou l'autre de ces modifications change la valeur p_c à laquelle la transition de phase se produit, la valeur des exposants critiques, quant à elle, ne dépend que de la dimension de la grille [42]. Ainsi, par exemple, toutes les grilles en deux dimensions se comportent identiquement autour de leur point critique et ce, indépendamment du détail de leurs configurations respectives.

Ceci étant dit, ces différentes variantes partagent toutes une même caractéristique fondamentale : la structure sur laquelle la percolation se produit, autrement dit la grille, est régulière et est, de ce fait, déterministe. Autrement dit, la nature probabiliste de ces modèles n'est due qu'à la présence aléatoire des sites et/ou des liens. En outre, la composante extensive occupe la totalité de la grille lorsque tous les sites et/ou les liens sont occupés. Or, qu'en est-il si la structure même de ces grilles est également aléatoire ? Dans cette thèse, nous répondons à cette question en étudiant la percolation sur des graphes aléatoires (ceux-ci sont définis aux sections suivantes). D'emblée, nous pouvons dire que l'émergence d'un amas extensif sera limitée par l'existence d'une *composante* extensive dans la structure des graphes aléatoires

[1] Une quantité extensive est une quantité qui varie linéairement avec la taille du système. L'amas extensif dont il est question contiendra donc deux fois plus de cases si la taille de la grille, N, est doublée.

[2] À défaut d'avoir trouvé une revue exhaustive récente, nous pointons le lecteur vers la page Wikipédia sur les seuils de percolation (`http://en.wikipedia.org/wiki/Percolation_threshold`, en date du 27 février 2014). Elle permet de constater la très grande diversité des modèles de percolation proposés au fil des années et pointe vers les articles qui les ont étudiés en détails.

[3] Bien qu'on présente généralement la théorie de la percolation par un modèle de percolation par sites, la percolation par liens fut étudiée en premier [31, 186, 187]. C'est d'ailleurs dans ce contexte que le nom « percolation » fut introduit : ces modèles étaient utilisés pour étudier l'écoulement d'un fluide dans un milieu poreux irrégulier. Les liens représentaient alors les chemins que pouvait emprunter le fluide.

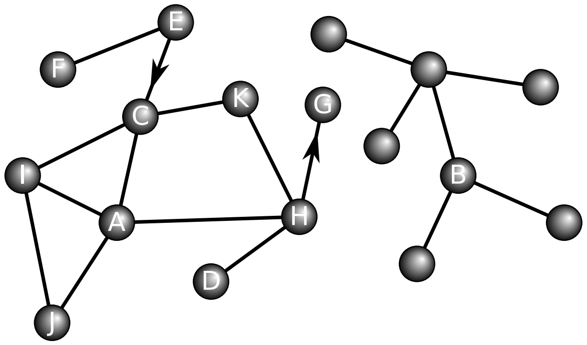


FIGURE 1.1 – Exemple simple d'un graphe utilisé pour illustrer la terminologie et les propriétés introduites à la section 1.1.

sous-jacents. De plus, comme nous le montrons à la section 1.3.2 de même qu'aux chapitres 3 et 4, la caractérisation de la structure de ces graphes aléatoires (ex. tailles des composantes, émergence d'une composante extensive) et l'étude de la percolation par sites ou par liens sur ceux-ci sont mathématiquement identiques. En fait, cette caractérisation correspond simplement à l'étude de la percolation lorsque tous les nœuds et/ou tous les liens existent. Ainsi, nous utiliserons la nomenclature de la théorie de la percolation également pour parler de la caractérisation de la structure des graphes aléatoires.

Afin de tout de même faire le pont entre la percolation sur graphes aléatoires et la percolation sur grilles régulières, mentionnons que les graphes aléatoires considérés dans cette thèse possèdent très peu de parcours fermés (c.-à-d. des boucles ou des cycles). Par conséquent, ces graphes occupent des espaces géométriques dont la dimensionnalité est très élevée, et sont en quelque sorte analogues aux grilles de Bethe. Ainsi, leur structure autour du point critique est décrite par les exposants critiques associés aux grilles dont la dimension est supérieure ou égale à 6 [42] : l'émergence de la composante extensive et la divergence de la taille moyenne des petites composantes sont toutes deux décrites par des exposants critiques égaux à 1.

1.1.2 Qu'est-ce qu'un graphe ?

Un graphe est un objet mathématique abstrait composé de deux unités fondamentales : les *nœuds* et les *liens* (*nodes* et *edges*). Lorsque utilisés dans un contexte de modélisation les nœuds représentent les éléments en interaction et les liens représentent les interactions. La figure 1.1 offre une illustration simple de ce qu'est un graphe. En se référant à cette figure, nous introduisons les termes suivants.

Voisins—Deux nœuds sont voisins s'ils sont joints par un lien. Nous dirons, lorsque nécessaire, qu'ils sont *premiers* voisins lorsqu'un seul lien les sépare (ils sont à une distance $d = 1$). Par extension, deux nœuds sont d-ième voisins si le plus petit parcours entre eux est composé de d liens. Par exemple, les nœuds A et C sont voisins, alors que les nœuds A et D sont seconds voisins.

Liens unidirectionnels et bidirectionnels—Il est possible que l'interaction entre deux nœuds soit inégale, le cas le plus extrême étant lorsque celle-ci est unidirectionnelle (ex. une page web peut contenir un lien vers une autre page sans que cela soit réciproque). Il est donc possible que certains liens d'un graphe ne puissent être parcourus que dans un seul sens. Le cas échéant, les deux nœuds sont joints à l'aide d'un *lien unidirectionnel* (*directed edge*). Par exemple, les nœuds C et E de même que les nœuds G et H sont joints par un lien unidirectionnel alors que tous les autres nœuds sont joints à l'aide de *liens bidirectionnels* (*undirected edges*).

Degré—Le degré est le nombre de premiers voisins qu'un nœud donné possède. Ainsi, le nœud A à la figure 1.1 a un degré égal à quatre. Lorsqu'un graphe possède des liens unidirectionnels, on distingue les degrés *entrant* et *quittant* (*incoming degree* et *outgoing degree*). Par exemple, en plus d'un degré bidirectionnel égal à trois, le nœud H possède un degré entrant nul et un degré quittant égal à un.

Degré sortant—Le degré sortant (*excess degree*) d'un nœud est le nombre de liens par lesquels il est possible de quitter un nœud sans repasser par le lien via lequel ce nœud a préalablement été atteint. Le degré sortant est en fait égal au degré d'un nœud moins un (le lien par lequel il a été atteint). Par exemple, le nœud B dans la figure 1.1 possède un degré sortant égal à deux.

Motif—Un motif est un groupe de nœuds, dont les liens sont disposés selon une configuration particulière, correspondant à une structure récurrente dans un graphe (ex. triangle, losange). Par exemple, dans les graphes extraits de systèmes complexes réels, on retrouve un nombre beaucoup plus élevé de triangles que dans un graphe équivalent dans lequel les liens ont été distribués aléatoirement (voir l'effet d'agrégation à la sous-section suivante).

Composante—Une composante est un groupe de nœuds isolés du reste du graphe où chacun des membres est le voisin (premier, second, etc.) des autres nœuds du groupe. Le graphe montré à la figure 1.1 possède deux composantes, dont une, celle de droite, est un *arbre*. On qualifiera de *petite* toute composante pour laquelle le ratio $s/N \to 0$ dans la limite $N \to \infty$ où N est le nombre de nœuds dans le graphe et s est le nombre de nœuds dans la composante (aussi appelé sa taille). Par opposition, on qualifiera de *géante* la composante pour laquelle le ratio $s/N \to \mathcal{S}$ dans la même limite avec $0 < \mathcal{S} \leq 1$. Autrement dit, le nombre de nœuds dans la composante géante est une quantité extensive. Sauf dans le cas particulier investigué au chapitre 5, la composante géante, lorsqu'elle existe, est unique. Lorsque certains nœuds d'une composante sont joints par des liens unidirectionnels, celle-ci peut alors être divisée en trois parties : *entrante*, *sortante* et *centrale* (*in-component*, *out-component* et *strongly connected component*) [32]. La partie centrale de la composante est constituée de l'ensemble maximal de nœuds dans lequel il existe un parcours entre chaque paire de nœuds possible (ex. tous les nœuds de la composante de gauche dans la figure 1.1 à l'exception des nœuds E, F et G). La partie entrante de la composante est constituée des nœuds à partir desquels il est possible d'atteindre la partie centrale, mais qui ne peuvent être atteints à partir de celle-ci (ex. les nœuds E et F). La partie sortante est quant à elle composée des nœuds pouvant être atteints à partir de la composante centrale sans que celle-ci leur soit accessible (ex. le nœud G).

Avant de passer aux propriétés « universelles » qu'on retrouve dans plusieurs systèmes complexes réels, définissons un dernier objet mathématique qui nous sera utile à quelques reprises dans les prochaines sections et chapitres.

Matrice d'adjacence—Il existe plusieurs façons d'encoder mathématiquement la structure d'un graphe (c.-à-d. quels sont les voisins des nœuds) et l'une d'entre elles est la *matrice d'adjacence* **A** (*adjacency matrix*). Il s'agit d'une matrice $N \times N$, où N est le nombre de nœuds dans le graphe, dont l'élément A_{ij} correspond au nombre de liens quittant le i-ième nœud vers le j-ième nœud. Par exemple, si ces nœuds sont joints par trois liens dont deux sont bidirectionnels et un est unidirectionnel du i-ième nœud au j-ième nœud, nous aurons $A_{ij} = 3$ et $A_{ji} = 2$. Ainsi, la matrice d'adjacence n'est typiquement pas symétrique sauf lorsque tous les liens sont bidirectionnels. Notons que chaque lien reliant un nœud à lui-même contribue pour 2 à l'élément diagonal correspondant. Autrement dit, la matrice **A** compte les demi-liens dans le graphe de sorte que, lorsque $\mathbf{A} = \mathbf{A}^{\mathrm{T}}$ (c.-à-d. tous les liens sont bidirectionnels), nous retrouvons le degré des nœuds

$$k_i = \sum_{j=1}^{N} A_{ij} \ , \tag{1.1}$$

et par conséquent

$$\sum_{i,j=1}^{N} A_{ij} = \sum_{i=1}^{N} k_i = 2M \ , \tag{1.2}$$

où M est le nombre total de liens dans le graphe.

1.1.3 Propriétés universelles des systèmes réels

Les études empiriques effectuées sur des graphes extraits de systèmes complexes réels ont révélé plusieurs propriétés *universelles* au sens où on les retrouve dans un grand nombre de systèmes qui n'ont à première vue aucun point en commun et dont la provenance est très variée (ex. citations d'articles scientifiques, *toile* alimentaire des écosystèmes[4], liens amicaux dans une population, réseau de distribution électrique). Nous présentons sommairement[5]

[4] Les relations prédateurs-proies n'ont effectivement rien d'une chaîne…

[5] Le lecteur intéressé pourra se référer à l'excellent ouvrage *Networks : An introduction* de Mark Newman [144] pour une revue récente et exhaustive.

quelques-unes de ces propriétés auxquelles nous ferons référence dans les prochaines sections et les prochains chapitres. Il s'agit en fait des propriétés les plus fondamentales qui devront ultimement être reproduites par tout modèle théorique se voulant réaliste. La présente thèse rend d'ailleurs compte des efforts faits en ce sens.

Distribution des degrés asymétrique—La distribution du nombre de voisins qu'ont les nœuds d'une très grande majorité de systèmes réels s'apparente, du moins à partir d'un certain degré $k_{\min}$, à une *loi de puissance*

$$P(k) \propto k^{-\tau} , \tag{1.3}$$

où $P(k)$ est la probabilité qu'un nœud ait un degré égal à k. Indépendamment de la forme mathématique précise[6], la propriété importante de ces distributions est qu'elles sont très asymétriques (voir la figure 1.2). Ainsi, dans plusieurs systèmes réels une fraction petite mais non négligeable des nœuds ont un degré très élevé alors que le degré des autres nœuds est faible. Ces *super-connecteurs* (*hubs*) jouent un rôle important dans la connectivité (c.-à-d. la résilience) des graphes [4, 147], leur contrôle et monitoring [111, 200] de même que dans tous processus de propagation ayant cours sur un graphe [27, 136, 157].

Effet *Small-World*—Dans la plupart des graphes extraits de systèmes réels, tout nœud peut être rejoint à partir de n'importe quel autre nœud (s'ils font partie de la même composante) en passant par seulement quelques nœuds intermédiaires [197]. Mathématiquement, on dira que la distance moyenne entre deux nœuds est de l'ordre de $\log N$ où N est le nombre de nœuds dans le graphe [147]. En d'autres termes, malgré le fait que le graphe soit de très grande taille, il suffira généralement que de quelques *sauts* pour atteindre n'importe quel autre nœud du graphe. Dans la vie de tous les jours, cet effet est traduit par l'exclamation : « Que le monde est petit ! ».

Effet d'agrégation—L'effet d'agrégation (*clustering*) correspond à la propension qu'ont des nœuds voisins à avoir plusieurs voisins en commun. Cet effet est présent dans plusieurs systèmes réels (voir le tableau 1.1), notamment dans les interactions humaines. Il est en quelque sorte analogue à l'adage : « les amis de mes amis sont mes amis ». Bien qu'il ne soit pas limité aux voisins immédiats[7], on mesure l'effet d'agrégation dans un graphe par la présence de triangles (trois nœuds rejoints par trois liens). Ainsi, notant $N_{\triangle}$ le nombre de triangles dans un graphe et $N_{\wedge}$ le nombre de parcours de longueur deux (c.-à-d. un triangle potentiel), on définit le coefficient d'agrégation comme étant

$$C = \frac{3N_{\triangle}}{N_{\wedge}} \tag{1.4}$$

[6] Un débat eut lieu pendant plusieurs années à savoir si les distributions des degrés étaient bel et bien toutes des lois de puissance pures [43, 140], ou si ces distributions appartenaient plutôt à différentes familles (ex. distribution log-normale). Le débat s'est quelque peu estompé depuis quelques années ; la communauté ayant consensuellement adopté les nuances appropriées.

[7] Nous pourrions en effet nous intéresser à la probabilité que deux nœuds qui ne sont pas voisins aient plus d'un voisin en commun (c.-à-d. aux cycles composés de quatre nœuds).

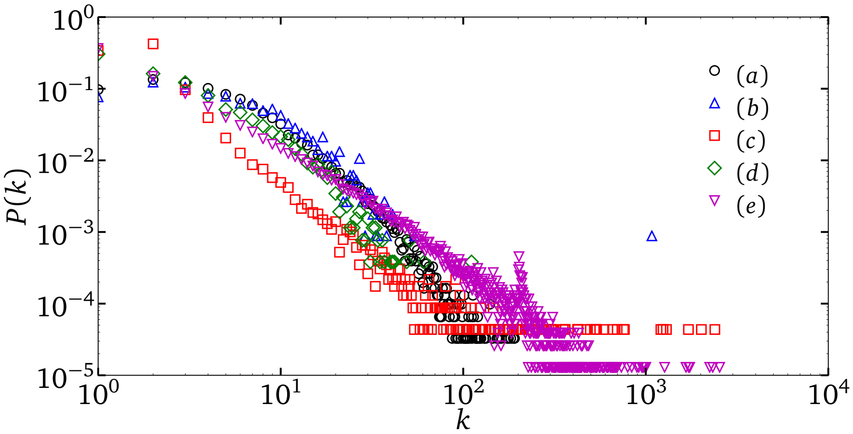


FIGURE 1.2 – Distributions des degrés de graphes extraits de systèmes réels. (a) Collaborations scientifiques : les auteurs (nœuds) sont joints s'ils ont coécrit au moins un article présent sur `arXiv.org` [26]. (b) Échanges de courriels : deux personnes sont jointes si elles ont échangé au moins un courriel [83]. (c) Extrait du réseau Internet (téléchargé du site web de M. E. J. Newman (`http://www-personal.umich.edu/~mejn/netdata`). (d) Graphe des interactions protéines-protéines de la levure *S. cerevisiae* dans lequel les protéïnes sont jointes si elles interagissent dans un quelconque processus biochimique [153]. (e) Réseau social des utilisateurs du site de nouvelles technologiques `slashdot.org` [110].

où le facteur trois au numérateur tient compte du fait qu'un triangle contient trois parcours de longueur deux. Ainsi, ce coefficient est borné par l'intervalle $[0, 1]$ et peut être interprété comme étant la probabilité que deux voisins d'un nœud soient également voisins. Cette quantité peut aussi être vue comme une mesure de la redondance des voisinages de nœuds voisins, de même que comme une mesure de la déviation d'un graphe d'une structure en arbre[8]. La figure 1.3 montre un sous-ensemble du graphe de la figure 1.1 pour lequel on compte $N_\triangle = 2$, $N_\wedge = 17$ et donc un coefficient d'agrégation $C = 6/17$. Il existe d'autres mesures pour quantifier l'effet d'agrégation. Citons notamment le coefficient local d'agrégation du i-ième nœud

$$c_i = \frac{\Delta_i}{\binom{k_i}{2}} = \frac{2\Delta_i}{k_i(k_i - 1)} \,, \tag{1.5}$$

où Δ_i est le nombre de triangles auxquels participe le i-ième nœud et où k_i est son degré. Ce coefficient est lui aussi borné par l'intervalle $[0, 1]$ et peut être interprété comme la probabilité que deux voisins du i-ième nœud soient également voisins. Par exemple, le nœud A à la figure 1.3 possède un coefficient local d'agrégation $c_A = 1/3$. Bien que l'ensemble des coefficients c_i offre une information locale très détaillée sur la structure d'un graphe, il est souvent plus révélateur de travailler avec des versions agglomérées de ceux-ci (ex. des

[8] Dans un arbre, deux nœuds voisins n'ont aucun voisin en commun. Comparons par exemple les deux composantes de la figure 1.1.

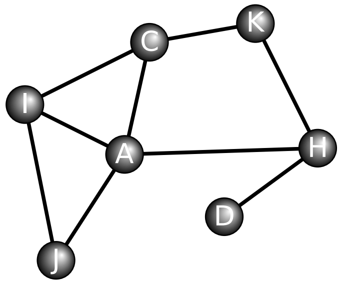


FIGURE 1.3 – Sous-ensemble du graphe montré à la figure 1.1 utilisé pour illustrer le calcul des coefficients d'agrégation.

distributions). Par exemple, en étudiant le coefficient local d'agrégation moyen en fonction du degré des nœuds,

$$\bar{c}(k) = \frac{2}{k(k-1)|\mathcal{Y}_k|} \sum_{i \in \mathcal{Y}_k} \Delta_i \ , \tag{1.6}$$

où $\mathcal{Y}_k$ est l'ensemble des nœuds de degré k et $|\mathcal{Y}_k|$ est sa cardinalité, il a été démontré que $\bar{c}(k) \propto k^{-\alpha}$ avec $0 < \alpha \leq 1$ pour plusieurs systèmes réels, et que ceci est possiblement une signature de leur organisation hiérarchique [162, 194]. Finalement, plutôt que d'utiliser le coefficient d'agrégation C, certains auteurs préfèrent travailler avec le coefficient local d'agrégation moyen [197]

$$\bar{c} = \frac{1}{N} \sum_{i=1}^{N} c_i = \frac{2}{N} \sum_{i=1}^{N} \frac{\Delta_i}{k_i(k_i-1)} \ . \tag{1.7}$$

Ce dernier s'interprète similairement au coefficient d'agrégation C, mais n'est en général pas égal à ce dernier[9]. Par exemple, pour le graphe de la figure 1.3, nous obtenons

$$\bar{c} = \frac{c_A + c_C + c_D + c_H + c_I + c_J + c_K}{7} = \frac{\frac{1}{3} + \frac{1}{3} + 0 + 0 + \frac{2}{3} + 1 + 0}{7} = \frac{1}{3} \ , \tag{1.8}$$

ce qui est légèrement inférieur à $C = 6/17$ tel que calculé précédemment (voir également le tableau 1.1). Nonobstant cette différence quantitative, ces deux coefficients possèdent le même comportement qualitatif de sorte que l'un ou l'autre peut être utilisé pour quantifier l'agrégation dans un graphe donné.

Structure communautaire—Amenant l'idée de redondance dans le voisinage des nœuds un peu plus loin, un très grand nombre d'études ont montré que les éléments constituant la plupart des systèmes complexes sont regroupés en *communautés* ou *modules*, c.-à-d. en groupe de nœuds qui sont plus densément joints les uns aux autres qu'avec les autres nœuds du graphe [2, 69, 79, 141]. En d'autres termes, contrairement à une composante (c.-à-d. un groupe *fermé*),

[9] La différence entre les deux provient du fait que C est le « ratio des moyennes » et que $\bar{c}$ est la « moyenne des ratios ».

Système complexe	N	$\langle k\rangle$	C	$\bar{c}$	r
Fréquentations estudiantines	573	1.67	0.005	0.001	-0.029
Réseau métabolique	765	9.64	0.09	0.67	-0.240
Réseau de distribution électrique	4941	2.67	0.10	0.08	-0.003
Internet	10697	5.98	0.035	0.39	-0.189
Collaboration scientifique (mathématiques)	253339	3.92	0.15	0.34	0.120
WWW (`nd.edu`)	269504	11.11	0.11	0.29	-0.067
Réseau d'acteurs (`IMDb.com`)	449913	113.43	0.20	0.78	0.208

TABLE 1.1 – Coefficients d'agrégation C et $\bar{c}$ de même que le coefficient de corrélation r pour sept graphes extraits de systèmes complexes (leur taille N de même que leur degré moyen $\langle k\rangle$ sont également donnés). Les données de ce tableau sont tirées de *Networks : An Introduction* de Mark E. J. Newman (tableau 8.1, page 237) [144].

une communauté est un groupe *ouvert* de nœuds partageant un grand nombre de premiers, seconds, etc. voisins. Ces modules peuvent, par exemple, représenter des groupes d'amis dans un réseau social ou des gènes co-régulés dans un micro-organisme. La présence de ces modules dans les graphes est généralement associée à une fonction commune ou collective des éléments correspondants dans le système complexe associé. C'est pourquoi beaucoup d'efforts ont été investis au cours des dernières années au développement d'algorithmes de détection qui tentent d'inférer ces communautés à partir de la structure d'un graphe. Les modèles présentés dans cette thèse n'ont pas comme objectif explicite de reproduire une structure communautaire réaliste. Comme nous l'expliquerons dans les prochains chapitres, l'approche théorique que nous avons développé offre cependant plusieurs voies intéressantes pour une telle modélisation (ex. corrélation via les types de nœuds, incorporation de *motifs* de grande taille), et c'est pourquoi nous mentionnons au passage le concept de structure communautaire.

Corrélations—L'effet d'agrégation et la structure communautaire mentionnés ci-dessus suggèrent qu'il peut exister des corrélations entre les attributs des deux nœuds aux extrémités des liens. Par exemple, le fait de savoir qu'un nœud fait partie d'un triangle implique que deux de ses voisins ont un degré au moins égal à deux. De même, les nœuds appartenant à une même communauté partagent fort probablement une certaine propriété, même si celle-ci n'est pas encodée explicitement dans la structure du graphe. Par exemple, pensons aux jeunes d'une ville qui risquent d'être davantage joints aux autres jeunes fréquentant la même école ou aux autres jeunes vivant dans le même quartier. En fait, ce genre de corrélations est bien connu en sociologie dans le cas des interactions humaines (voir par exemple [21, 131]), et plus récemment il fut montré que des corrélations similaires existent aussi dans des systèmes complexes de différentes natures [36, 135, 137, 144].

Étant donné que l'information est explicitement encodée dans le graphe, la corrélation entre le degré de nœuds voisins fut particulièrement étudiée. Il fut d'ailleurs démontré que pour beaucoup de systèmes complexes, la probabilité qu'un lien joigne deux nœuds de degrés k et k', $P(k,k')$, s'écarte significativement de la valeur attendue si les liens étaient attribués aléatoirement. Il s'agit d'une propriété importante puisque ce genre de corrélations, selon lesquelles les nœuds tendent à être joints à des nœuds de degré similaire ou différent, forge la structure globale du graphe et par conséquent influence toute dynamique qu'il sous-tend [145]. À l'instar du coefficient d'agrégation, il est parfois plus révélateur de travailler avec le coefficient de corrélation suivant

$$r = \frac{\langle kk' \rangle - \langle k \rangle \langle k' \rangle}{\langle k^2 \rangle - \langle k \rangle^2} \ , \tag{1.9}$$

où $\langle x \rangle \equiv \sum_{k,k'} xP(k,k')$ [56]. Par définition, ce coefficient est limité à l'intervalle $[-1,1]$. On aura $r < 0$ lorsque les liens ont tendance à joindre des nœuds de degrés différents, $r > 0$ lorsque ces nœuds ont tendance à avoir un degré similaire, et $r = 0$ en l'absence de corrélations. Fait intéressant, il semble que les systèmes complexes réels puissent être différenciés à l'aide du coefficient r. En effet, comme le suggère partiellement les données montrées au tableau 1.1, il semble que les interactions sociales (au sens large) soient caractérisées par des coefficients r positifs, alors que tous les autres types de systèmes complexes (ex. technologique, biologique) tendent plutôt à avoir un coefficient r négatif [119, 135, 146, 144, 156].

1.2 Graphes aléatoires de type Erdős-Rényi

Modèle phare de la théorie des graphes, les graphes aléatoires de type Erdős-Rényi sont un incontournable pour toute introduction à la percolation sur graphes et sont à la base des contributions présentées au chapitre 2 et au début du chapitre 4. Popularisés par Paul Erdős (1913–1996) et Alfred Rényi (1921–1970) qui en étudièrent les propriétés structurelles au tournant des années 1960 [62, 63, 64, 65], les graphes aléatoires de type Erdős-Rényi sont devenus un champ de recherche à part entière [29] et sont encore à ce jour l'objet de recherche active (voir par exemple [133]). De ce fait, un corpus considérable fut généré au fil des dernières décennies et nous nous limiterons, par souci de concision, qu'aux résultats en lien direct avec les travaux de cette thèse. Le lecteur intéressé pourra se tourner notamment vers l'aride *Random Graphs* de Béla Bollobás [29] pour une revue exhaustive des propriétés des graphes aléatoires.

1.2.1 Définition et propriétés élémentaires

À l'origine, Erdős et Rényi étudièrent en détails les propriétés structurelles de l'ensemble de graphes aléatoires $\mathcal{G}_{N,M}$. Les graphes de cet ensemble sont composées de N nœuds discernables et de M liens distribués aléatoirement entre les $\binom{N}{2}$ paires de nœuds possibles. Ainsi, $\mathcal{G}_{N,M}$ définit un ensemble de $\binom{\binom{N}{2}}{M}$ graphes dans lequel tous les graphes sont équiprobables.

À cet égard, l'ensemble $\mathcal{G}_{N,M}$ est analogue à l'ensemble micro-canonique en physique statistique où les états correspondant à un même état macroscopique—ici le nombre *fixe* de liens M, analogue à l'énergie—sont équiprobables. La figure 1.4 donne un exemple d'un graphe de l'ensemble $\mathcal{G}_{N,M}$.

Il est toutefois d'usage de considérer le pendant *canonique* de l'ensemble $\mathcal{G}_{N,M}$: l'ensemble $\mathcal{G}_{N,p}$ introduit par Eugene N. Gilbert (1923–2013) également au tournant des années 1960 [78]. Cet ensemble est composé de N nœuds discernables dans lesquels chacun des $\binom{N}{2}$ liens possibles existe indépendamment des autres liens avec une probabilité p. Ainsi, l'ensemble $G_{N,p}$ est composé de $2^{\binom{N}{2}}$ graphes, et chacun des $\binom{\binom{N}{2}}{m}$ graphes possédant m liens y apparaît avec une probabilité égale à

$$p^m(1-p)^{\binom{N}{2}-m} \,. \tag{1.10}$$

Cet ensemble est analogue à l'ensemble canonique en physique statistique ; le paramètre p jouant le rôle de la température en fixant le nombre moyen de liens, $\langle m \rangle = \binom{N}{2}p$, analogue à l'énergie moyenne. Étant donné l'indépendance de l'existence d'un lien par rapport à celle des autres, la probabilité $P_{\mathrm{ER}}(k)$ qu'un nœud ait un degré égal à k est distribuée de façon binomiale :

$$P_{\mathrm{ER}}(k) = \binom{N-1}{k} p^k (1-p)^{N-1-k} \,, \tag{1.11}$$

de laquelle on calcule aisément le degré moyen des nœuds

$$z \equiv \langle k \rangle = \sum_{k=0}^{N-1} k P_{\mathrm{ER}}(k) = (N-1)p \,. \tag{1.12}$$

À l'exception des résultats faisant l'objet de la section 1.2.3, nous nous intéressons en général à la structure des graphes des ensembles $\mathcal{G}_{N,M}$ et $\mathcal{G}_{N,p}$ dans la limite $N \to \infty$. Plus particulièrement, les résultats obtenus par Erdős-Rényi qui nous intéressent sont obtenus lorsque $M \sim cN$ où c est un nombre réel positif (incluant zéro)[10] [11]. Cette relation implique que le degré moyen des nœuds dans l'ensemble $\mathcal{G}_{N,M}$ tend vers une constante lorsque N tend vers l'infini

$$\langle k \rangle = \frac{2M}{N} \sim 2c \,, \tag{1.13}$$

où le facteur 2 présent au numérateur tient compte du fait qu'un lien contribue au degré de deux nœuds. Or, le nombre de liens, m, dans les graphes de l'ensemble $G_{N,p}$ est distribué

[10] Formellement, ce nombre est borné supérieurement par $c = (N-1)/2$ de sorte que $M = \binom{N}{2}$, ce qui correspond au cas où les nœuds partagent un lien avec tous les autres nœuds du graphe. Puisque nous nous intéressons à la limite $N \to \infty$, cette borne *s'éloigne* à l'infini de sorte que c peut prendre n'importe quelle valeur dans l'intervalle $[0, \infty)$.

[11] Ici la notation $M \sim cN$ indique simplement que le ratio $\frac{M}{cN} \to 1$ lorsque $N \to \infty$.

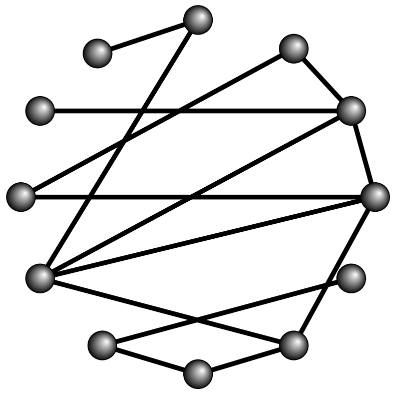


FIGURE 1.4 – Exemple d'un graphe aléatoire de type Erdős-Rényi avec $N = 12$ et $M = 14$. Ce graphe particulier se produit dans les ensembles $\mathcal{G}_{N,M}$ et $\mathcal{G}_{N,p}$ avec une probabilité égale à $\binom{\binom{N}{2}}{M}^{-1}$ et à $p^M(1-p)^{\binom{N}{2}-M}$ respectivement.

selon

$$\binom{\binom{N}{2}}{m} p^m (1-p)^{\binom{N}{2}-m} \, , \tag{1.14}$$

qui est une distribution très piquée autour de sa valeur moyenne

$$\langle m \rangle = \sum_{m=0}^{\binom{N}{2}} m \binom{\binom{N}{2}}{m} p^m (1-p)^{\binom{N}{2}-m} = p\binom{N}{2} = \frac{zN}{2} \tag{1.15}$$

lorsque $N \to \infty$. Ainsi, en choisissant $\langle m \rangle = M$, on peut s'attendre à ce que les deux ensembles soient peuplés de façon similaire dans cette limite, et par conséquent que plusieurs de leurs propriétés asymptotiques soient les mêmes. En plus de suggérer une possible équivalence entre les deux ensembles, cette analyse offre, le cas échéant, une traduction, $z = 2c$, entre les paramètres z et c (et donc entre p et M).

De fait, sous certaines conditions, les ensembles $G_{N,p}$ et $G_{N,M}$ peuvent être utilisés de façon interchangeable [29, 159] ; les paramètres p et M étant reliés via $p = M/\binom{N}{2}$. Les conditions rigoureuses permettant cette équivalence s'éloignant quelque peu du cadre de cette thèse[12], nous nous contenterons de mentionner que celle-ci requiert que $p(1-p)\binom{N}{2} \to \infty$, une condition qui est respectée lorsque le degré moyen z est constant et borné (la situation d'intérêt dans le cadre de cette thèse), à savoir,

$$p(1-p)\binom{N}{2} = \frac{z}{2}N - \frac{z^2}{2}\frac{N}{N-1} \sim \frac{z}{2}N \to \infty \, , \tag{1.16}$$

[12] Le lecteur intéressé peut se référer chapitre 2.1 de l'ouvrage *Random Graphs* de Béla Bollobás [29].

où nous avons utilisé $z = (N-1)p$. Comme nous le verrons tout au long de cette section, l'indépendance d'existence des liens dans la formulation de l'ensemble $\mathcal{G}_{N,p}$ en facilite grandement le traitement mathématique. L'équivalence entre les deux ensembles offrant en quelque sorte le choix de travailler avec l'un ou l'autre des ensembles, l'ensemble $\mathcal{G}_{N,p}$ est généralement choisi au détriment l'ensemble $\mathcal{G}_{N,M}$. Ainsi, quoique historiquement erronée, l'appellation « graphes aléatoires de type Erdős-Rényi » fait aujourd'hui généralement référence à l'ensemble $\mathcal{G}_{N,p}$ et non à l'ensemble $\mathcal{G}_{N,M}$. Cette thèse souscrit à cette appellation.

Remarques sur la distribution des degrés

Puisque que nous nous intéressons aux propriétés structurelles des graphes aléatoires de type Erdős-Rényi dans la limite $N \to \infty$ mais où le degré moyen des nœuds est constant et borné, il est bon de mentionner que dans ces conditions la distribution des degrés de l'ensemble $\mathcal{G}_{N,p}$ [équation (1.11)] tend vers une distribution de Poisson[13] de moyenne z,

$$P_{\mathrm{ER}}(k) = \binom{N-1}{k} p^k (1-p)^{N-1-k} \sim \frac{z^k \mathrm{e}^{-z}}{k!} \ , \tag{1.17}$$

d'où l'utilisation de l'appellation alternative *graphes de Poisson* pour l'ensemble $\mathcal{G}_{N,p}$.

1.2.2 Structure globale et transition de phase

Le grand intérêt à l'égard les graphes aléatoires de type Erdős-Rényi provient principalement du comportement spectaculaire de la plus grande composante présente dans les graphes de l'ensemble en fonction de la densité de liens (exprimée à l'aide de z). En effet, si $s_{\max}(N)$ est le nombre de nœuds dans cette composante, Erdős et Rényi démontrèrent qu'à la densité critique $z = 1$, cette composante cesse d'occuper une fraction négligeable des nœuds du graphe ($s_{\max}(N)/N \to 0$ lorsque $z < 1$) et devient extensive ($s_{\max}(N)/N \sim \mathcal{S}$ où $0 < \mathcal{S} \leq 1$ lorsque $z > 1$) [63].

Pour les raisons expliquées à la section 1.2.4, les travaux effectués dans le cadre de cette thèse se basent sur un ensemble plus réaliste de graphes aléatoires (c.-à-d. *le modèle des configurations*) et seuls les résultats présentés à la section 1.2.3 seront utilisés dans les chapitres

[13] Ce résultat se démontre en considérant la fonction génératrice de la distribution (1.11),

$$\sum_{k=0}^{\infty} P_{\mathrm{ER}}(k) x^k = \sum_{k=0}^{N-1} \left[\binom{N-1}{k} p^k (1-p)^{N-1-k} \right] x^k = [1 + p(x-1)]^{N-1} \ ,$$

où la seconde égalité tient en vertu du théorème binomial. Dans la limite $N \to \infty$,mais où $z = p(N-1)$ est borné, nous obtenons

$$\lim_{N\to\infty} [1 + p(x-1)]^{N-1} = \lim_{N\to\infty} \left[1 + \frac{z(x-1)}{N-1} \right]^{N-1} = \mathrm{e}^{z(x-1)} = \sum_{k=0}^{\infty} \left[\frac{z^k \mathrm{e}^{-z}}{k!} \right] x^k \ ,$$

ce qui correspond à la fonction génératrice de la distribution de Poisson. Il s'agit bien entendu d'un résultat connu. Nous avons toutefois profité de l'occasion pour illustrer l'utilité des fonctions génératrices qui seront abondamment utilisées tout au long de cette thèse.

suivants. Nous jugeons toutefois pertinent de présenter, ne serait-ce que sommairement, les résultats de Erdős et de Rényi, et ce pour deux raisons principales. La première relève de considérations esthétiques : ces résultats sont beaux. Nous ne donnerons que les résultats principaux dans ce chapitre, et par conséquent nous ne rendrons que partiellement justice à l'élégance de l'analyse de Erdős et de Rényi. Nous invitons donc le lecteur à consulter la référence [63] pour plus de détails. La seconde justification relève de considérations pratiques : bien que nous ne travaillerons pas directement avec l'ensemble $\mathcal{G}_{N,p}$, les résultats de Erdős et de Rényi restent qualitativement valides pour le modèle des configurations. De plus, la simplicité de l'ensemble $\mathcal{G}_{N,p}$ permet d'acquérir une bonne compréhension de la mécanique sous-jacente à l'évolution de sa structure en fonction de la densité de liens, et cette compréhension peut directement se transposer au modèle des configurations.

Régime sous-critique ($z < 1$)

Lorsque le degré moyen des nœuds est inférieur à 1 ($z < 1$), il n'y a *presque sûrement* [14] pas de composante extensive dans l'ensemble $\mathcal{G}_{N,p}$. La question est donc de savoir de quelle façon les nœuds et liens sont organisés ; la réponse à cette question sera primordiale à l'élaboration du formalisme présenté à la section 1.3.

Erdős et de Rényi démontrèrent que lorsque $M \sim cN$, ce qui est équivalent à fixer le degré moyen z dans $\mathcal{G}_{N,p}$, les nœuds sont presque sûrement tous regroupés dans des composantes en arbre ou dans des composantes ne contenant qu'un seul cycle. Le nombre de nœuds dans ces composantes étant distribué selon une distribution de Poisson dont la moyenne dépend du type de composantes (ex. arbre, cycle). Plus précisément, ils démontrèrent que le nombre moyen de nœuds appartenant à des cycles, $\langle N_{\mathrm{c}}(N) \rangle$, converge vers une constante

$$\langle N_{\mathrm{c}}(N) \rangle \sim \frac{z^3}{2(1-z)} \, , \tag{1.18}$$

et que le nombre moyen de nœuds appartenant à des composantes en arbre, $\langle N_{\mathrm{a}}(N) \rangle$, converge vers

$$\langle N_{\mathrm{a}}(N) \rangle \sim N \, . \tag{1.19}$$

Autrement dit, les composantes dans le régime sous-critique sont essentiellement des arbres ; il s'agit-là de la propriété qui sera centrale à la section 1.3. Ils démontrèrent enfin que la plus grande composante est un arbre (les composantes contenant des cycles étant de taille

[14] La nature probabiliste de l'ensemble $\mathcal{G}_{N,p}$ implique qu'il existe toujours un graphe contredisant un résultat énoncé dans cette section. Par exemple, même lorsque $z < 1$, il existe un graphe où tous les liens existent. Celui-ci se produit toutefois avec une probabilité $p^{\binom{N}{2}} = [z/(N-1)]^{\binom{N}{2}}$ ce qui devient rapidement négligeable dans la limite $N \to \infty$. Par conséquent, nous utilisons l'expression *presque sûrement* (traduction de *almost surely*) pour signifier une situation dans laquelle un résultat se produit avec une probabilité de 1, mais où il existe tout de même un ensemble de mesure zéro (c.-à-d. de probabilité nulle) de graphes contredisant ledit résultat.

modérée) et que[15]

$$s_{\max}(N) = O(\log N) \;, \tag{1.20}$$

presque sûrement, c'est-à-dire qu'aucune composante dans le régime sous-critique n'a une taille extensive.

Transition de phase $(z = 1)$ et régime surcritique $(z > 1)$

L'équation (1.18) offre un aperçu du changement structurel de $\mathcal{G}_{N,p}$ à l'approche du seuil $z = 1$: le nombre de nœuds dans les composantes contenant des cycles s'emballe. Ceci est symptomatique du fait que le nombre de nœuds dans des composantes contenant qu'un seul cycle, $N_{\mathrm{c},1}(N)$, augmente significativement lors de la transition de phase

$$N_{\mathrm{c},1}(N) \sim N^{2/3} \;. \tag{1.21}$$

Il en va de même pour la taille de la plus grande composante (toujours un arbre) qui occupe alors une fraction plus importante du graphe

$$s_{\max}(N) \sim N^{2/3} \;, \tag{1.22}$$

mais dont la taille n'est toutefois toujours pas extensive.

Lorsque le degré moyen excède ce point critique ($z > 1$), la plus grande composante devient extensive et sa taille est telle que

$$s_{\max}(N) \sim \mathcal{S}N \;, \tag{1.23}$$

où $0 < \mathcal{S} \leq 1$ est la solution de

$$\mathcal{S} = 1 - \mathrm{e}^{-z\mathcal{S}} \;. \tag{1.24}$$

Cette composante, dite *géante*, est unique et ne possède pas une structure en arbre. En contrepartie, la fraction $1 - \mathcal{S}$ des nœuds qui ne font pas partie de cette composante géante sont répartis dans des composantes dont la structure est similaire à celle des composantes dans le régime sous-critique. Ces composantes sont essentiellement des arbres et leur taille est bornée à $O(\log N)$.

Remarques sur la taille de la composante géante

Il est intéressant de constater que la transition de phase correspondant à l'émergence de la composante géante est encodée dans l'équation (1.24). En effet, si on solutionne cette équation lorsque $z < 1$, on ne trouve que la solution $\mathcal{S} = 0$ dans l'intervalle $[0, 1]$, ce qui correspond bel et bien au fait qu'il n'y a pas de composante géante dans le régime sous-critique.

[15] La notation $O(\cdots)$ indique une borne supérieure dans la limite $N \to \infty$. Dans ce cas d'espèce, nous avons $s_{\max}(N) = O(\log N)$ ce qui signifie que $\frac{s_{\max}(N)}{\log N}$ est non nul et borné.

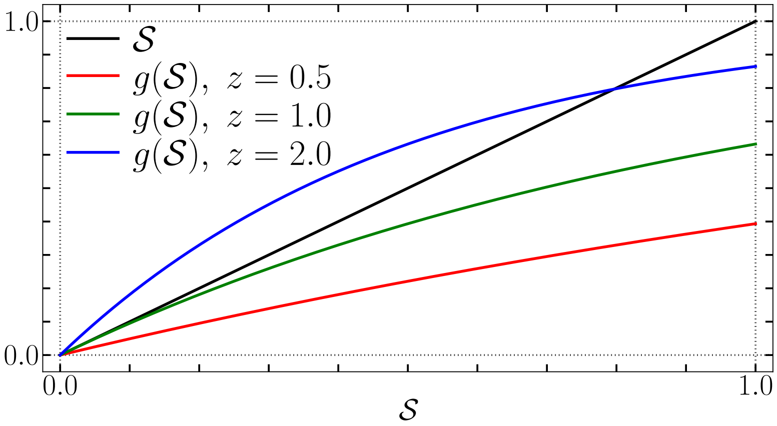


FIGURE 1.5 – Analyse graphique de l'équation (1.24). Si le côté droit de l'équation, la fonction $g(\mathcal{S})$, possède une pente supérieure à 1 à $\mathcal{S} = 0$, c.-à-d. si $g'(0) = z > 1$, l'équation (1.24) aura une seconde solution dans l'intervalle $[0, 1]$ correspondant à la présence d'une composante géante dans l'ensemble $\mathcal{G}_{N,p}$. À l'opposé, si la pente à $\mathcal{S} = 0$ est inférieure à 1, c.-à-d. si $g'(0) = z < 1$, la seule solution dans $[0, 1]$ est $\mathcal{S} = 0$, ce qui correspond au régime sous-critique et à l'absence d'une composante extensive.

Or, lorsque $z > 1$, une deuxième solution $\mathcal{S} > 0$ correspondant à la taille relative de la composante géante apparaît dans $[0, 1]$.

Ce comportement s'explique directement en analysant graphiquement[16] l'équation (1.24) (voir la figure 1.5). En effet, nous voyons que le côté droit de l'équation, que nous notons $g(\mathcal{S})$, est une fonction croissante et concave

$$g'(\mathcal{S}) = \frac{\partial g(\mathcal{S})}{\partial \mathcal{S}} = z\mathrm{e}^{-z\mathcal{S}} > 0 \tag{1.25}$$

$$g''(\mathcal{S}) = \frac{\partial^2 g(\mathcal{S})}{\partial \mathcal{S}^2} = -z^2\mathrm{e}^{-z\mathcal{S}} < 0 \tag{1.26}$$

pour toute valeur de $\mathcal{S}$ puisque $z > 0$. De plus, puisque z est borné, nous savons que

$$g(\mathcal{S}) = 1 - \mathrm{e}^{-z\mathcal{S}} \leq 1\,, \tag{1.27}$$

où l'égalité ne tient que lorsque $\mathcal{S} \to \infty$. Par conséquent, puisque $g(0) = 0$, nous voyons que si $g'(0) = z < 1$, les deux côtés de l'équation (1.24) ne coïncideront jamais pour tout $\mathcal{S} > 0$. A contrario, si $g'(0) = z > 1$, les deux côtés de l'équation (1.24) se recroiseront dans l'intervalle $0 < \mathcal{S} \leq 1$ étant donnée l'inégalité (1.27). Ces conclusions sont illustrées à la figure 1.5 et la taille de la composante géante en fonction du degré moyen z est montrée à la figure 1.6.

[16] Une analyse plus complète en termes de la convergence d'un système dynamique itératif associé est présentée à la section 1.3.3.

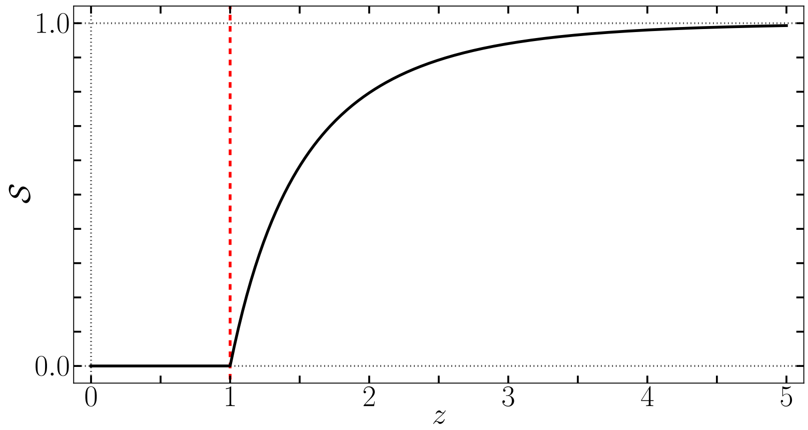


FIGURE 1.6 – Taille de la composante géante dans l'ensemble $\mathcal{G}_{N,p}$ en fonction du degré moyen z. La courbe noire représente la plus grande solution de l'équation (1.24) dans l'intervalle $[0, 1]$ et la ligne pointillée rouge représente la valeur du seuil ($z = 1$) séparant les régimes sous-critique ($z < 1$) et surcritique ($z > 1$).

1.2.3 Résultats sur des graphes de taille finie

La simplicité de l'ensemble $\mathcal{G}_{N,p}$ permet de calculer exactement la distribution de la taille des composantes dans des graphes dont la taille n'excède pas quelques centaines, voire quelques milliers de nœuds [78, 101, 138]. Cette distribution s'obtient en itérant une condition initiale simple à l'aide de deux équations qui sont définies ci-dessous. Ces équations sont le point de départ des contributions présentées au chapitre 2 et au début du chapitre 4.

Définissons $W(n|N, p)$ comme la probabilité qu'un nœud quelconque dans un graphe de l'ensemble $\mathcal{G}_{N,p}$ soit dans une composante contenant n nœuds. Pour obtenir $W(n|N, p)$, considérons un graphe de l'ensemble $\mathcal{G}_{N,p}$ contenant une composante de taille n dont fait partie le nœud A. Afin que cette composante soit isolée, il est nécessaire que chacun des n nœuds de la composante soient individuellement isolés des $N - n$ autres nœuds du graphe. En d'autres termes, aucun des $n(N - n)$ liens reliant les membres de ces deux groupes ne doit exister, ce qui se produit avec une probabilité $(1 - p)^{n(N-n)}$. Puisqu'il y a $\binom{N-1}{n-1}$ façons de choisir les $n - 1$ nœuds du graphe faisant partie de la même composante que A et que ceux-ci forment une composante connectée avec une probabilité $W(n|n, p)$, nous obtenons la première relation de récurrence

$$W(n|N, p) = \binom{N-1}{n-1}(1 - p)^{n(N-n)}W(n|n, p) . \tag{1.28}$$

En d'autres termes, sachant que le nœud A fait partie d'une composante de n nœuds, cette équation permet en quelque sorte d'ajouter des nœuds au graphe tout en préservant la taille de la composante à laquelle appartient le nœud A. Notons que la composante de taille n n'est pas restreinte aux n nœuds « originaux » ; une fois les $N - n$ nœuds ajoutés au graphe,

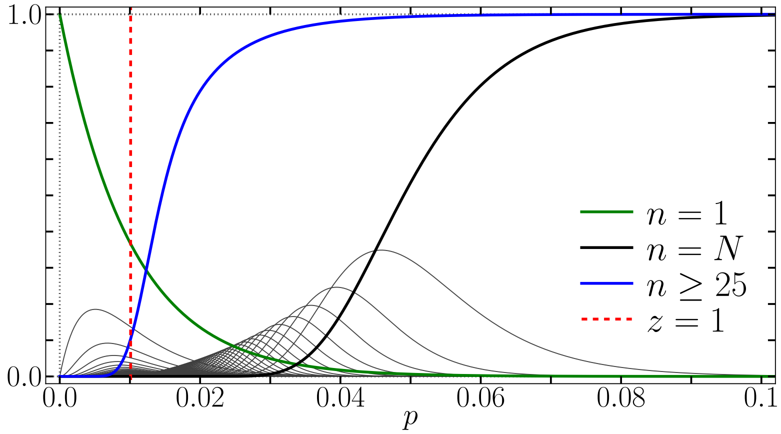


FIGURE 1.7 – Distribution de la taille des composantes, $\{W(n|N,p)\}_{n=1,\ldots,N}$, dans l'ensemble $\mathcal{G}_{N,p}$ en fonction de la probabilité d'existence des liens p pour $N = 100$. Chaque courbe grise représente le comportement de $P(n|N,p)$ pour une taille n donnée ; l'ordre (de gauche à droite) dans lequel ces courbes culminent correspond aux tailles de 2 à 99 en ordre croissant. Les courbes pour les tailles $n = 1$ et $n = N = 100$ sont tracées en vert et en noir. À titre indicatif, nous avons marquée la valeur de p correspondant à $z = 1$ [$p = 1/(N-1)$] et avons également tracé (courbe bleue) la courbe correspondant à la probabilité que la composante ait une taille supérieure ou égale à 25.

tous les nœuds peuvent faire partie de cette composante, d'où le terme $\binom{N-1}{n-1}$ dans l'équation (1.28). Pour obtenir $W(n|n,p)$, il suffit de noter que pour une taille de graphe donnée, les probabilités $W(n|N,p)$ sont normalisées, c.-à-d. $\sum_{l=1}^{n} W(l|n,p) = 1$. Par conséquent, nous avons

$$W(n|n,p) = 1 - \sum_{l=1}^{n-1} W(l|n,p) \tag{1.29}$$

où les termes du côté droit de l'équation sont préalablement obtenus à l'aide de l'équation (1.28) pour des graphes dans lesquels le nœud A appartient à une composante de taille l. Ultimement, ce calcul itératif requiert la condition initiale $W(1|1,p) = 1$ qui stipule simplement que la probabilité de trouver une composante de taille 1 dans un graphe contenant un seul nœud est égale à 1.

La figure 1.7 montre le comportement des probabilités $W(n|N,p)$ en fonction de p pour $N = 100$. Il est intéressant de constater que, même à cette petite taille, on perçoit déjà qualitativement le comportement de l'ensemble $\mathcal{G}_{N,p}$ dans la limite $N \to \infty$. La figure 1.7 montre en effet l'omniprésence des composantes de petite taille pour $z < 1$, de même que la probabilité croissante d'une composante occupant une grande proportion du graphe pour $z > 1$. Ajoutons également qu'on y voit déjà la dichotomie dans la taille des composantes observée dans la limite $N \to \infty$: la probabilité de trouver une composante de taille intermédiaire est très faible (c.-à-d. soit les nœuds font partie de composantes de petite taille ou ils font partie d'une composante de grande taille [17]). À titre indicatif, nous avons ajouté une courbe (en bleue) montrant la probabilité d'y trouver une composante contenant au moins 25 nœuds (taille légèrement supérieure à $N^{2/3}$ choisie arbitrairement).

Le calcul de la distribution $\{W(n|N,p)\}_{n=1,\ldots,N}$ pour une taille N et une probabilité p données nécessite le calcul d'environ $\binom{N}{2}$ probabilités (dont certaines sont presque nulles) pour les tailles intermédiaires. Par conséquent, quoique qu'exactes, l'utilisation des équations (1.28) et (1.29) est en pratique limitée à des graphes de quelques centaines, voire quelques milliers de nœuds.

1.2.4 L'ensemble $\mathcal{G}_{N,p}$ en tant que modèle de systèmes réels

L'ensemble $\mathcal{G}_{N,p}$ est l'un des modèles les plus simples possédant une transition de phase « percolative ». Comme nous venons de le présenter, cette simplicité permet l'obtention de résultats analytiques, et la nature épurée de l'ensemble permet de développer une bonne intuition sur le mécanisme menant à l'émergence de la composante géante. Cette simplicité a toutefois son revers et en fait un piètre modèle pour reproduire les propriétés des graphes tirés de systèmes complexes réels. En effet, à l'exception de l'effet Small-World (présent dans la très grande majorité des graphes aléatoires), l'ensemble $\mathcal{G}_{N,p}$, dans lequel la distribution des degrés est binomiale, ne reproduit pas les distributions des degrés hautement asymétriques (voir la figure 1.2) trouvées dans la plupart des systèmes réels [15, 140]. De plus, on n'y retrouve aucune agrégation puisque que la probabilité que deux voisins d'un nœud soient également voisins est simplement égale à la probabilité qu'un lien existe entre ces deux nœuds, autrement dit $C = p = z/(N-1) = O(N^{-1})$, une probabilité qui tend vers zéro dans la limite $N \to \infty$. Il n'existe aucune corrélation dans la façon dont les nœuds sont connectés les uns aux autres, et on y retrouve aucune structure communautaire. Par conséquent, l'ensemble $\mathcal{G}_{N,p}$ est un mauvais candidat pour toute tentative de modélisation de systèmes réels.

1.3 Modèle des configurations

Parmi les propriétés structurelles des systèmes réels qui ne sont pas reproduites par les graphes aléatoires de type Erdős-Rényi, l'une des plus importantes est la distribution des degrés. Tel que mentionné à la section 1.1.3, la distribution des degrés possède une influence déterminante et prépondérante sur la structure globale des graphes (ex. connectivité, robustesse) de même que sur plusieurs processus dynamiques ayant cours sur ceux-ci (ex. propagation d'agents contagieux, contrôle et surveillance de systèmes complexes). Par conséquent, un premier pas logique vers un modèle théorique de la structure des interactions

[17] Notons toutefois qu'à $N = 100$, plusieurs composantes de grandes tailles peuvent encore coexister.

sous-jacentes des systèmes complexes réels est l'incorporation de distributions des degrés réalistes dans un modèle de graphes aléatoires.

Le *modèle des configurations* (*Configuration Model*[18]) est le candidat majoritairement choisi pour effectuer l'incorporation de distributions des degrés réalistes. Plusieurs raisons motivent cette préférence. Premièrement, ce modèle est conceptuellement simple et permet de spécifier une distribution des degrés arbitraire. Deuxièmement, il est possible de calculer exactement plusieurs propriétés structurelles à l'aide de *fonctions génératrices de probabilités* (voir la section 1.3.2). Troisièmement, sa simplicité conceptuelle combinée à la souplesse de l'approche par fonctions génératrices permet facilement de généraliser le modèle pour y incorporer une multitude d'autres propriétés structurelles présentes dans les graphes extraits de systèmes complexes réels. Puisqu'une part importante de cette thèse rend compte des efforts déployés pour généraliser le modèle des configurations, nous consacrons le reste de ce chapitre à l'analyse détaillée de ce modèle[19].

1.3.1 Définition et propriétés élémentaires

Soit N nœuds discernables dont le degré est prescrit par la séquence $\{k_i\}_{i=1,\ldots,N}$ dans laquelle le j-ième terme, k_j, correspond au degré du j-ième nœud. Tel qu'illustré à la figure 1.8a, la séquence $\{k_i\}_{i=1,\ldots,N}$ peut être vue comme étant N nœuds desquels émergent un nombre de *demi-liens* (*stubs*) égal à leur degré respectif. Le modèle des configurations s'intéresse à l'ensemble des graphes pouvant être générés à partir de cette séquence en jumelant aléatoirement par paire tous les demi-liens pour former les liens entre les nœuds[20]. Nous appellerons cet ensemble de graphes l'*ensemble CM*. Conséquemment, la somme des degrés

$$\sum_{i=1}^{N} k_i = N\langle k\rangle = 2M \, , \tag{1.30}$$

où $\langle k\rangle$ est le degré moyen et M est le nombre total de liens, doit être pair afin que tous les demi-liens puissent être jumelés. Notons que cette procédure n'exclût pas que deux demi-liens appartenant au même nœud soient jumelés, ni que deux nœuds soient joints par plus

[18] Ce modèle fut introduit en 1978 par Edward A. Bender and E. Rodney Canfield [22] puis étudié plus en profondeur notamment par Béla Bollobás [28, 29]. Les auteurs s'intéressèrent entre autres au nombre de graphes pouvant être générés à partir d'une séquence de degrés (voir la section 1.3.1) dans la limite $N \to \infty$. Étonnamment, il fallut attendre les deux articles phares de Michael Molloy et Bruce Reed au milieu des années 1990 [129, 130] pour que la transition de phase du modèle soit rigoureusement définie et investiguée. La formulation à l'aide des fonctions génératrices (voir la section 1.3.2), de même que son utilisation en tant que modèle pour des systèmes complexes réels sont en très grande partie dues à Mark E. J. Newman [139, 144, 147] qui sortit le modèle de son giron purement mathématique et l'introduisit à la physique et à la science des réseaux complexes.

[19] La présentation du modèle des configurations faite dans cette section est fortement inspirée par celle faite par Mark E. J. Newman dans *Networks : An Introduction* [144]. Nous avons toutefois adapté la notation par souci d'uniformité avec celle utilisée dans les prochains chapitres, et certaines parties, notamment l'analyse des solutions de l'équation (1.64) en termes d'une bifurcation transcritique, est originale.

[20] Algorithmiquement, on crée une liste dans laquelle le nœud i apparaît k_i fois (c.-à-d. chaque demi-lien apparaît une fois). On mélange ensuite cette liste et on jumelle les $(2n-1)$-ième et $(2n)$-ième éléments.

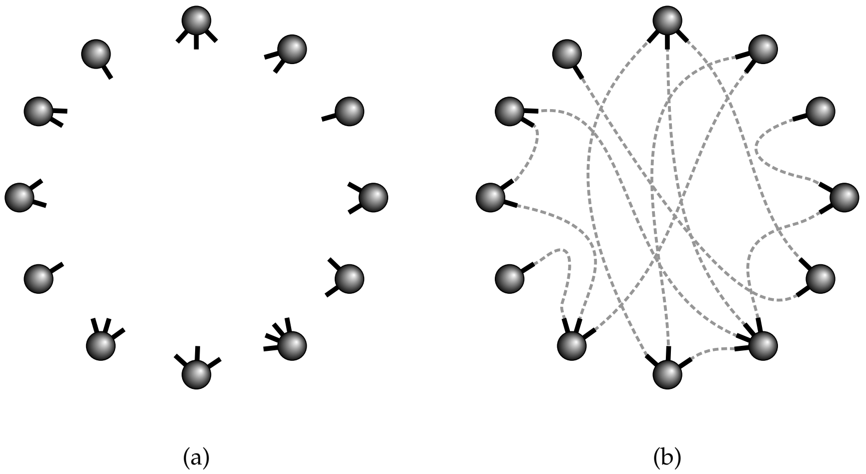


FIGURE 1.8 – Illustration de la construction des graphes de l'ensemble CM. (a) Chacun des $N = 12$ nœuds possède un nombre de demi-liens égal à son degré. Il y a trois nœuds de degré 1, cinq nœuds de degré 2, trois nœuds de degrés 3 et un nœud de degré 4. (b) Les graphes de l'ensemble sont obtenus en joignant aléatoirement les demi-liens.

d'un lien. Toutefois, bien qu'elles ne soient pas souhaitables en général, laisser ces « anomalies structurelles » se produire permet de préserver la simplicité du modèle en évitant de lui imposer des contraintes supplémentaires. Qui plus est, comme nous le montrerons plus loin dans cette section, ces anomalies n'ont pas d'impact sur les propriétés structurelles globales des graphes générés par le modèle dans la limite $N \to \infty$.

Chaque exécution de la procédure décrite au paragraphe précédant génère un *arrangement* particulier des demi-liens (ex. le troisième demi-lien du i-ième nœud a été jumelé au deuxième demi-lien du j-ième nœud, etc.), et il est facile de se convaincre que chaque arrangement est équiprobable. Cependant, puisque nous ne nous intéressons qu'aux graphes où seuls les nœuds sont discernables, plusieurs arrangements correspondent à un même graphe[21]. Plus précisément, un graphe dont la matrice d'adjacence est $\mathbf{A} = \{A_{ij}\}_{i,j=1,\ldots,N}$ peut être obtenue à partir de

$$\frac{\prod_i k_i!}{\prod_{i<j} A_{ij}! \prod_i A_{ii}!!} \tag{1.31}$$

arrangements différents [rappelons que $\sum_{j=1}^{N} A_{ij} = k_i$, que $n!! = n(n-2)(n-4)\ldots$ et que $0! = 1$, par convention]. Le produit au numérateur est simplement le nombre total de façons dont l'ordre des demi-liens peut être permuté tout en préservant la structure du graphe. Toutefois, lorsqu'il existe plus d'un lien entre deux nœuds ou lorsqu'un nœud est joint à lui-

[21] Le nom du modèle provient du fait que chaque graphe pouvant être généré à partir d'une séquence de degrés est aussi appelé une *configuration*.

même, certaines de ces permutations correspondent au même arrangement (ex. la permutation des deux demi-liens formant un lien entre un nœud et lui-même) et doivent donc être retirées du compte (rôle des deux produits au dénominateur). Par conséquent, les graphes possédant plus d'un lien entre deux nœuds ou des liens liant des nœuds à eux-mêmes seront sous-représentés par rapport aux graphes exempts de ces anomalies structurelles. Cependant, ces anomalies sont extrêmement rares lorsque $N \to \infty$, et par conséquent la procédure de création de graphes décrites précédemment peut être considérée, de façon effective, comme échantillonnant uniformément l'ensemble CM.

Fréquence des anomalies structurelles

Il est possible d'estimer le nombre moyen de paires de nœuds joints par plus d'un lien de même que le nombre moyen de liens liant un nœud à lui-même dans la limite $N \to \infty$. Notons d'abord que puisque les demi-liens sont jumelés aléatoirement, la probabilité que les i-ième et j-ième nœuds soient joints est

$$\frac{k_i k_j}{2M-1} \simeq \frac{k_i k_j}{2M} \,. \tag{1.32}$$

Ceci correspond au ratio entre le nombre de façons dont les i-ième et j-ième nœuds peuvent être joints, soit $k_i k_j$, et le nombre total de façons de jumeler l'ensemble des $2M$ demi-liens, soit $\binom{2M}{2} = M(2M-1)$, le tout multiplié par le nombre de liens qui seront créés ultimement (c.-à-d. le nombre total de *chances* pour créer un lien, soit M). De façon similaire, s'il existe déjà un lien entre les i-ième et j-ième nœuds, la probabilité qu'un second lien existe entre ceux-ci est

$$\frac{(k_i-1)(k_j-1)}{2M} \,, \tag{1.33}$$

de sorte que la probabilité qu'il y ait au moins deux liens entre les i-ième et j-ième nœuds est

$$\frac{k_i k_j (k_i-1)(k_j-1)}{(2M)^2} \,. \tag{1.34}$$

En sommant sur toutes les paires de nœuds possibles, nous obtenons une estimation du nombre moyen de paires de nœuds joints par plus d'un lien

$$\begin{aligned}\sum_{i=1}^{N}\sum_{j=i+1}^{N} \frac{k_i k_j (k_i-1)(k_j-1)}{(2M)^2} &\simeq \frac{1}{2(2M)^2}\left[\sum_{i=1}^{N} k_i(k_i-1)\right]\left[\sum_{j=1}^{N} k_j(k_j-1)\right] \\ &= \frac{1}{2\langle k\rangle^2}\left[\frac{1}{N}\sum_{i=1}^{N} k_i(k_i-1)\right]^2 \\ &= \frac{1}{2}\left[\frac{\langle k^2\rangle - \langle k\rangle}{\langle k\rangle}\right]^2 \,,\end{aligned} \tag{1.35}$$

où nous avons utilisé l'équation (1.30) de même que $\langle k^2\rangle = \frac{1}{N}\sum_{i=1}^{N} k_i^2$. Pour ce qui est du nombre moyen de liens liant un nœud à lui-même, notons que la quantité équivalente à l'expression (1.32) s'obtient en multipliant le ratio entre le nombre de façons de choisir deux demi-liens du i-ième nœud $[\binom{k_i}{2} = k_i(k_i-1)/2]$, et le nombre total de façons de jumeler l'ensemble des $2M$ demi-liens $[\binom{2M}{2} = M(2M-1)]$, par le nombre de liens qui seront créés ultimement, M :

$$\frac{k_i(k_i-1)}{2(2M-1)} \simeq \frac{k_i(k_i-1)}{4M} \; . \tag{1.36}$$

En sommant sur tous les nœuds du graphe, nous obtenons que le nombre moyen de liens liant un nœud à lui-même est environ

$$\sum_{i=1}^{N} \frac{k_i(k_i-1)}{4M} = \frac{1}{2}\left[\frac{\langle k^2\rangle - \langle k\rangle}{\langle k\rangle}\right] \; . \tag{1.37}$$

Les expressions (1.32) et (1.37) indiquent donc que le nombre moyen de paires de nœuds joints par plus d'un lien et le nombre moyen de liens liant un nœud à lui-même tendent tous les deux vers des constantes lorsque $N \to \infty$ et que $\langle k^2\rangle$ est borné. Par conséquent, leur densité respective tend vers zéro dans cette même limite, ce qui nous permet de négliger la présence de ces anomalies lors des calculs effectués dans le reste de cette section.

Structure globale et transition de phase

À l'instar des graphes aléatoires de type Erdős-Rényi, les graphes de l'ensemble CM subissent une transition de phase correspondant à l'émergence d'une composante géante (c.-à-d. une composante contenant un nombre extensif de nœuds) dans la limite $N \to \infty$ avec $\langle k\rangle$ borné. En fait, Molloy et Reed démontrèrent que la structure globale des graphes de l'ensemble CM est qualitativement très similaire à celle de l'ensemble $\mathcal{G}_{N,p}$ [129, 130]. En effet, lorsque

$$\frac{\langle k^2\rangle - \langle k\rangle}{\langle k\rangle} < 1 \; , \tag{1.38}$$

il n'y a pas de composante géante, les composantes sont essentiellement des arbres, et la taille de la plus grande d'entre-elles est

$$s_{\max}(N) = O(k_{\max}^2 \log N) \; , \tag{1.39}$$

presque sûrement, où $k_{\max}$ est le plus haut degré dans la séquence $\{k_i\}_{i=1,\ldots,N}$. Au contraire, lorsque

$$\frac{\langle k^2\rangle - \langle k\rangle}{\langle k\rangle} > 1 \; , \tag{1.40}$$

il existe presque sûrement une composante géante dont la taille relative est donnée par les équations (1.64) et (1.69). Les autres composantes sont essentiellement des arbres et leur taille est bornée à $O(\log N)$.

Ensemble de graphes défini via la distribution des degrés

Il arrive fréquemment que nous soyons davantage intéressés à étudier un ensemble de graphes défini via une distribution des degrés $\{P(k)\}_{k\in\mathbb{N}}$ donnant la probabilité $P(k)$ qu'un nœud ait un degré égal à k plutôt que via une séquence $\{k_i\}_{i=1,\ldots,N}$. Pour un nombre de nœuds N donné, on s'intéresse alors à l'ensemble des séquences $\{k_i\}_{i=1,\ldots,N}$ pouvant être générées en pigeant aléatoirement et indépendamment N degrés de la distribution $\{P(k)\}_{k\in\mathbb{N}}$. Chaque séquence apparaît donc avec une probabilité $\prod_{i=1}^{N} P(k_i)$ dans cet ensemble. Pour toute observable $X(\{k_i\})$ calculée pour l'ensemble de graphes défini par une de ces séquences, la valeur, $\langle X\rangle$, de cette même observable pour l'ensemble de graphes défini par la distribution des degré $\{P(k)\}_{k\in\mathbb{N}}$ est simplement la moyenne des $X(\{k_i\})$ pondérées par la probabilité de la séquence associée :

$$\langle X\rangle = \sum_{k_1=0}^{\infty} \ldots \sum_{k_N=0}^{\infty} X(\{k_i\}) \prod_{i=1}^{N} P(k_i)\,. \tag{1.41}$$

Puisque nous nous intéressons aux propriétés du modèle des configurations dans la limite $N \to \infty$, la loi des grands nombres [67] nous assure que les proportions de chaque degré dans les séquences générées converge vers la distribution des degrés $\{P(k)\}_{k\in\mathbb{N}}$. Par conséquent, à l'instar des ensembles $\mathcal{G}_{N,M}$ et $\mathcal{G}_{N,p}$ définis à la section 1.2, les deux approches sont équivalentes dans la limite $N \to \infty$. Nous utiliserons ce fait dans les prochaines sections.

1.3.2 Fonctions génératrices de probabilité

Nous aurons recours aux fonctions génératrices de probabilité pour étudier l'organisation structurelle (c.-à-d. présence d'une composante extensive, distribution de la taille des petites composantes) des graphes de l'ensemble CM dans la limite $N \to \infty$. Ce choix est motivé par le fait que ces fonctions possèdent plusieurs propriétés qui permettent un traitement mathématique exact et intuitif. Nous présenterons ces diverses propriétés au fur et à mesure que nous introduirons les fonctions génératrices nécessaires aux calculs effectués dans le reste de ce chapitre. Notons que, bien que ces propriétés seront illustrées à l'aide d'une fonction particulière, celles-ci tiennent cependant pour toutes les autres fonctions génératrices utilisées dans cette thèse sauf si le contraire est explicitement mentionné.

Une fonction génératrice de probabilité est simplement une série de puissances dont les coefficients représentent une distribution de probabilité. Par exemple, la première fonction génératrice qui nous est nécessaire est celle associée à la distribution des degrés $\{P(k)\}_{k\in\mathbb{N}}$

$$g(x) = \sum_{k=0}^{\infty} P(k)x^k = P(0) + P(1)x + P(2)x^2 + P(3)x^3 + \ldots \tag{1.42}$$

Autrement dit, le coefficient devant le terme x^n est la probabilité qu'un nœud ait un degré égal à n. Notons d'abord que puisque ces fonctions sous-tendent des distributions de probabilité, nous avons que

$$g(1) = \sum_{k=0}^{\infty} P(k) = 1 \; , \tag{1.43}$$

à moins que le contraire soit mentionné explicitement (voir par exemple la section 1.3.4). Notons également qu'il est possible de retrouver chaque coefficient, $P(k)$, individuellement à partir d'une fonction génératrice par différentiation

$$P(k) = \left[\frac{1}{k!}\frac{\partial^k}{\partial x^k} g(x)\right]_{x=0} \; . \tag{1.44}$$

Il existe donc une relation bijective entre une distribution de probabilité et la fonction génératrice associée. Ceci nous permet donc d'utiliser ces deux quantités de façon interchangeable selon que l'une ou l'autre forme facilite le calcul. Les fonctions génératrices de probabilité permettent également le calcul des différents moments (lorsqu'ils existent) de la distribution qu'elles sous-tendent par différentiation. En effet, le m-ième moment s'obtient de la manière suivante :

$$\langle k^m \rangle = \left[\frac{\partial^m g(x)}{\partial (\ln x)^m}\right]_{x=1} = \Bigg[\underbrace{\left(x\frac{\partial}{\partial x}\right)\cdots\left(x\frac{\partial}{\partial x}\right)}_{m\text{ fois}} g(x)\Bigg]_{x=1} = \sum_{k=0}^{\infty} k^m P(k) \; . \tag{1.45}$$

En particulier, le degré moyen des nœuds de l'ensemble CM s'obtient en dérivant $g(x)$ une fois :

$$\langle k \rangle = g'(1) = \sum_{k=0}^{\infty} kP(k) \; , \tag{1.46}$$

où le prime dénote la dérivée première d'une fonction par rapport à son argument.

Supposons maintenant que nous dressions une liste de tous les liens dans un graphe de l'ensemble CM en notant le degré des nœuds aux extrémités (ex. le lien #3 joint un nœud de degré 3 et un nœud de degré 5). Les nœuds dont le degré est k apparaissent $NkP(k)$ fois dans cette liste puisque chacun de ces nœuds possède k demi-liens. Ainsi, le degré d'un nœud à l'une ou l'autre des extrémités d'un lien choisi au hasard dans cette liste est égal à k avec une probabilité

$$\frac{NkP(k)}{N\sum_{k'=0}^{\infty} k'P(k')} = \frac{kP(k)}{\langle k \rangle} \; . \tag{1.47}$$

Tel que vu à la section 1.3.1, les graphes de l'ensemble CM sont construits en jumelant aléatoirement les demi-liens émergeant des nœuds. Ceci a pour effet qu'il n'existe aucune corrélation entre le degré des nœuds aux deux extrémités des liens à l'exception de la corrélation « naturelle » donnée par l'équation (1.47). Ainsi, la probabilité $P(k,k')$ qu'un lien choisi aléatoirement joigne des nœuds de degré k et k' est

$$P(k,k') = \frac{kk'P(k)P(k')}{\langle k \rangle^2} \; . \tag{1.48}$$

La seconde fonction génératrice qui nous est nécessaire pour étudier l'organisation structurelle des graphes de l'ensemble CM génère la distribution des degrés *sortants* des nœuds (*excess degree*). Rappelons que le degré sortant est le nombre de liens par lequel il est possible de quitter un nœud atteint via un de ses liens (c.-à-d. en ne revenant pas sur nos pas). Or, étant donné la corrélation « naturelle » mentionnée ci-haut, un nœud n'aura pas un degré sortant égal à k_s avec une probabilité $P(k_\mathrm{s})$ (c.-à-d. la distribution des degrés), mais plutôt avec une probabilité égale à

$$P_\mathrm{s}(k_\mathrm{s}) = \sum_{k'=0}^{\infty} P(k_\mathrm{s}+1, k') = \frac{(k_\mathrm{s}+1)P(k_\mathrm{s}+1)}{\langle k \rangle} \ , \tag{1.49}$$

où le « +1 » provient du fait que le degré sortant d'un nœud est égal à son degré moins un (c.-à-d. le lien par lequel il a été atteint). La seconde fonction génératrice qui nous sera utile est donc

$$\begin{aligned} f(x) &= \sum_{k_\mathrm{s}=0}^{\infty} P_\mathrm{s}(k_\mathrm{s}) x^{k_\mathrm{s}} \\ &= \frac{1}{\langle k \rangle} \sum_{k_\mathrm{s}=0}^{\infty} (k_\mathrm{s}+1)P(k_\mathrm{s}+1) x^{k_\mathrm{s}} \\ &= \frac{1}{\langle k \rangle} \sum_{k=0}^{\infty} kP(k) x^{k-1} \\ &= \frac{g'(x)}{g'(1)} \ . \end{aligned} \tag{1.50}$$

Cette nouvelle fonction génératrice en main, nous nous intéressons à la distribution du nombre de *seconds* voisins, c.-à-d. le nombre de nœuds joignables en suivant successivement deux liens. Étant donné la structure en arbre des composantes dans la limite $N \to \infty$, un nœud dont le degré est k aura m seconds voisins avec une probabilité égale à

$$\sum_{m_1=0}^{\infty} \cdots \sum_{m_k=0}^{\infty} \delta\left(m, \sum_{i=1}^{k} m_i\right) \prod_{i=1}^{k} P_\mathrm{s}(m_i) \ , \tag{1.51}$$

où m_i est le degré sortant du i-ième voisin du nœud de degré k et où $\delta(m,n)$ est le delta de Kronecker. La probabilité que ce nœud ait un degré égal à k étant $P(k)$, la fonction générant la distribution du nombre de seconds voisins d'un nœud quelconque est

$$\begin{aligned} g^{(2)}(x) &\equiv \sum_{m=0}^{\infty} \left[\sum_{k=0}^{\infty} P(k) \sum_{m_1=0}^{\infty} \ldots \sum_{m_k=0}^{\infty} \delta\left(m, \sum_{i=1}^{k} m_i\right) \prod_{i=1}^{k} P_{\mathrm{s}}(m_i) \right] x^m \\ &= \sum_{k=0}^{\infty} P(k) \left[\sum_{m_1=0}^{\infty} \ldots \sum_{m_k=0}^{\infty} x^{\sum_{i=1}^{k} m_i} \prod_{i=1}^{k} P_{\mathrm{s}}(m_i) \right] \\ &= \sum_{k=0}^{\infty} P(k) \left[\sum_{i=0}^{\infty} P_{\mathrm{s}}(i) x^i \right]^m \\ &= \sum_{k=0}^{\infty} P(k) \left[f(x) \right]^m \\ &= g\big(f(x)\big) \; . \end{aligned} \tag{1.52}$$

Ce résultat met en lumière une autre propriété importante des fonctions génératrices de probabilité : leur produit correspond à la convolution des distributions qu'elles sous-tendent. Ainsi, la fonction $[f(x)]^m$ correspond à $m-1$ convolutions de la distribution des degrés sortants avec elle-même. Le coefficient devant le terme x^n dans $[f(x)]^m$ est donc la probabilité que m liens mènent à un total de n seconds voisins[22]. De l'équation (1.52), nous pouvons calculer le nombre moyen de seconds voisins qu'ont les nœuds de l'ensemble CM

$$\langle k \rangle_2 \equiv \left[\frac{\partial g^{(2)}(x)}{\partial x} \right]_{x=1} = g'(1) f'(1) = \langle k^2 \rangle - \langle k \rangle \; . \tag{1.53}$$

Par un même raisonnement, on peut montrer que la distribution du nombre de d-ièmes voisins est générée par

$$g^{(d)}(x) \equiv g\big(f\big(\ldots f(x) \ldots\big)\big) \; , \tag{1.54}$$

où la fonction $f(x)$ apparaît $d-1$ fois [147]. Le nombre moyen de d-ièmes voisins est alors

$$\langle k \rangle_d \equiv \left[\frac{\partial g^{(d)}(x)}{\partial x} \right]_{x=1} = \left(\frac{\langle k \rangle_2}{\langle k \rangle} \right)^{d-1} \langle k \rangle \; . \tag{1.55}$$

Nous voyons donc que le nombre de nœuds atteignables en s'éloignant d'un nœud quelconque croît exponentiellement lorsque

$$\frac{\langle k \rangle_2}{\langle k \rangle} = \frac{\langle k^2 \rangle - \langle k \rangle}{\langle k \rangle} > 1 \; , \tag{1.56}$$

ce qui correspond au critère pour l'existence d'une composante géante [voir l'équation (1.40)].

[22] Notons que la structure en arbre des composantes et l'absence de corrélation dans le jumelage des demi-liens sont primordiales pour assurer que les m « piges » du nombre de seconds voisins auxquels un lien mène soient indépendantes et donc que le nombre total de seconds voisins soit bel et bien le résultat d'une convolution.

Coefficient d'agrégation

Ayant défini la distribution des degrés sortants [équation (1.49)], nous pouvons calculer le coefficient d'agrégation dans les graphes de l'ensemble CM. Rappelons que ce coefficient correspond à la probabilité que deux voisins d'un même nœud soient également voisins. Autrement dit, il s'agit de la probabilité que « l'ami de mon ami soit aussi mon ami ». Cette quantité peut également être vue comme une mesure de la déviation de la structure des composantes d'une structure en arbre pure. Considérons le nœud i et deux de ses voisins u et v dont le degré est respectivement $k_u + 1$ et $k_v + 1$. Tel que vu à la section 1.3.1, la probabilité qu'un lien existe entre les nœuds u et v est approximativement $k_u k_v / 2M$ (un de leurs demi-liens est déjà jumelé à un demi-lien du nœud i). En pondérant cette probabilité par la probabilité que chacun de ces nœud ait un degré sortant égal à k_u et k_v (rappelons que ces nœuds ont été atteints via un de leurs liens), nous obtenons le coefficient d'agrégation

$$\begin{aligned} C &\equiv \sum_{k_u=0}^{\infty} \sum_{k_v=0}^{\infty} P_{\mathrm{s}}(k_u) P_{\mathrm{s}}(k_v) \frac{k_u k_v}{2M} \\ &= \frac{1}{2M} \left[\sum_{k=0}^{\infty} k P_{\mathrm{s}}(k) \right]^2 \\ &= \frac{1}{N \langle k \rangle^3} \left[\sum_{k=0}^{\infty} k(k+1) P(k+1) \right]^2 \\ &= \frac{1}{N} \frac{[\langle k^2 \rangle - \langle k \rangle]^2}{\langle k \rangle^3} \, , \end{aligned} \tag{1.57}$$

où nous avons utilisé l'équation (1.49) et le fait que $N\langle k \rangle = 2M$. Ainsi, pour un second moment de la distribution des degrés borné, nous concluons qu'il n'y a pas d'agrégation dans l'ensemble CM dans la limite $N \to \infty$. Ceci offre un autre indice pointant vers le fait que les petites composantes sont essentiellement des arbres.

Convexité des fonctions génératrices de probabilité

Puisque les fonctions génératrices de probabilité sont des séries de puissances dont tous les coefficients sont positifs (ce sont des probabilités), elles sont des fonctions croissantes et convexes dans l'intervalle d'intérêt $[0, 1]$. Prenons par exemple la fonction $g(x)$ pour laquelle nous obtenons clairement que

$$\frac{\partial g(x)}{\partial x} = \sum_{k=0}^{\infty} k P(k) x^{k-1} \geq 0 \, , \tag{1.58}$$

$$\frac{\partial^2 g(x)}{\partial x^2} = \sum_{k=0}^{\infty} k(k-1) P(k) x^{k-2} \geq 0 \, , \tag{1.59}$$

pour $0 \leq x \leq 1$. Les égalités se produisent à $x = 0$ si $P(1) = 0$ ou $P(2) = 0$, respectivement. La figure 1.9 illustre ces propriétés pour une distribution des degrés exponentielle (voir la section 1.3.5).

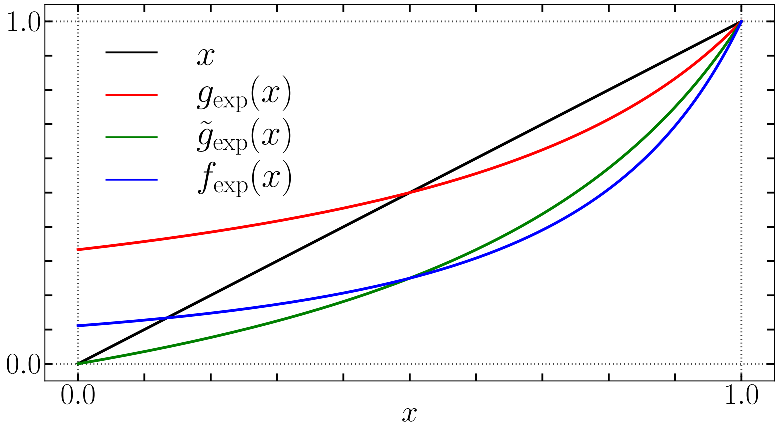


FIGURE 1.9 – Comportement typique des fonctions $g(x)$, $\tilde{g}(x)$ et $f(x)$ pour $x \in [0,1]$ pour une distribution des degrés exponentielle avec $\langle k \rangle = 2.0$ (voir section 1.3.5). Tel que démontré dans le texte, on y constate la convexité de ces fonctions de même que le fait que $\tilde{g}(x) \leq x$ dans ce même intervalle.

Remarque sur les nœuds de degré nul

Bien que leur existence soit admise par la définition de la distribution des degrés que nous utilisons, les nœuds de degré nul n'influencent aucunement la structure globale des graphes de l'ensemble CM puisque ceux-ci, n'ayant pas de liens, sont condamnés à rester en marge des petites composantes et d'une éventuelle composante extensive. Il est donc possible, en toute généralité, de retirer les nœuds de degré nul du graphe au besoin pour faciliter la manipulation mathématique des fonctions génératrices[23]. Mathématiquement, le retrait des nœuds de degré nul se fait via la substitution suivante

$$g(x) \to \tilde{g}(x) \equiv \frac{g(x) - g(0)}{1 - g(0)} \,, \tag{1.60}$$

qui lorsque appliquée à l'équation (1.50) et aux équations (1.64) et (1.69) ci-dessous nous permet de démontrer que le seul effet des nœuds de degré nul est de limiter la taille relative de la composante géante [c.-à-d. $\mathcal{S} \leq 1 - P(0)$].

Retrait aléatoire des nœuds et/ou des liens

Il arrive que nous nous intéressions aux propriétés structurelles des graphes de l'ensemble CM lorsque les nœuds et/ou les liens ont été retirés aléatoirement et indépendamment des graphes avec une probabilité $1 - r$ et $1 - p$ respectivement (c.-à-d. percolation par nœuds et/ou par liens). Étant donnée la structure effective en arbre des petites composantes dans la

[23] Par exemple, puisque $g(0) = P(0) = 0$ pour un ensemble de graphes exempt de nœuds de degré nul, on conclut que $g(x) \leq x$ dans tout l'intervalle $[0,1]$. Ceci est illustré à la figure 1.9 pour une distribution des degrés exponentielle (voir la section 1.3.5)

limite $N \to \infty$, ce retrait des nœuds ou des liens n'a pour effet que de modifier la distribution des degrés. Un nœud conservera donc son degré si le lien correspondant est toujours là et si le nœud à l'autre bout de ce lien existe toujours (probabilité rp). Puisque les nœuds et les liens sont retirés indépendamment les uns des autres, la nouvelle distribution des degrés est

$$\bar{P}(m) = \sum_{k=m}^{\infty} P(k) \binom{k}{m} (rp)^m (1-rp)^{k-m} \ , \tag{1.61}$$

où le terme binomial correspond à la probabilité qu'un nœud de degré k à l'origine conserve m voisins. La fonction génératrice associée à cette nouvelle distribution des degrés est

$$\begin{aligned} \bar{g}(x) &= \sum_{m=0}^{\infty} \bar{P}(m) x^m \\ &= \sum_{m=0}^{\infty} \sum_{k=m}^{\infty} P(k) \binom{k}{m} (rpx)^m (1-rp)^{k-m} \\ &= \sum_{k=0}^{\infty} P(k) \sum_{m=0}^{k} \binom{k}{m} (rpx)^m (1-rp)^{k-m} \\ &= \sum_{k=0}^{\infty} P(k) \big(1 - rp + rpx\big)^k \\ &= g\big(1 - rp + rpx\big) \ , \end{aligned} \tag{1.62}$$

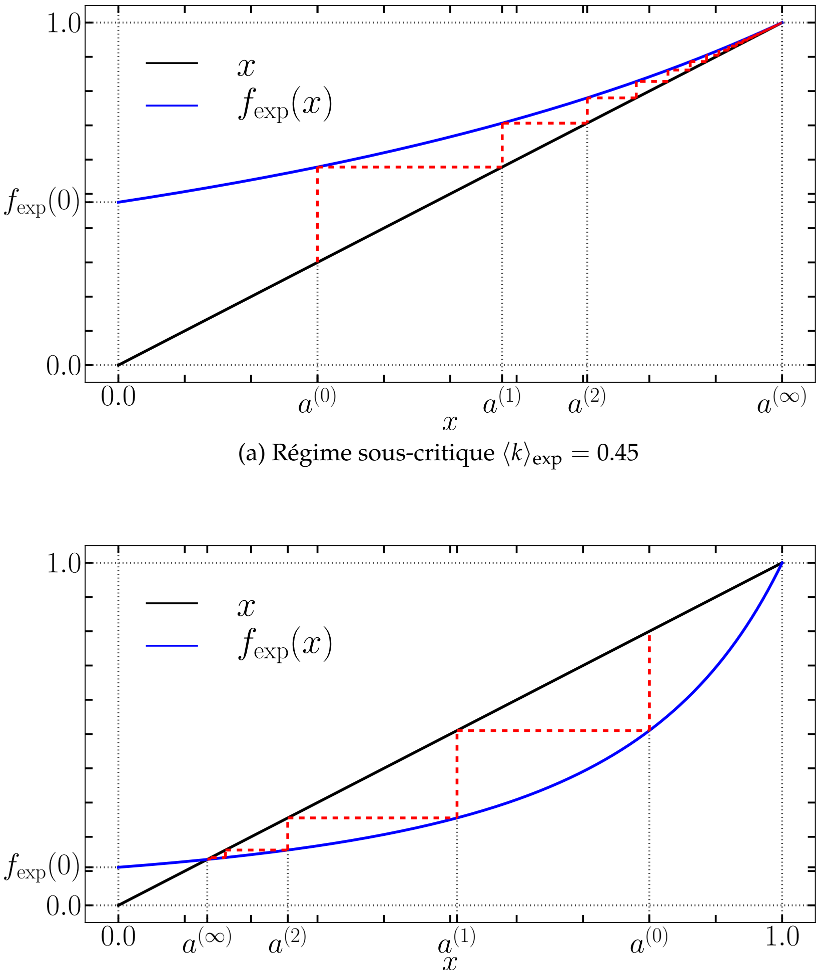
où l'inversion de l'ordre de sommation correspond à sommer les termes d'une matrice triangulaire rangée par rangée plutôt que colonne par colonne. L'équation (1.62) correspond donc à la même fonction génératrice, $g(x)$, que dans le cas où tous les nœuds et liens existent (c.-à-d. lorsque $r = p = 1$) pour laquelle nous avons fait la substitution

$$x \to 1 - rp + rpx \ . \tag{1.63}$$

Ainsi, tous les résultats obtenus pour l'ensemble CM s'appliquent également à la percolation sur cet ensemble en effectuant la substitution ci-dessus. Notons ici que les probabilités r et p apparaissent sous la forme rp ce qui laisse croire que la percolation par liens et la percolation par nœuds sont des concepts interchangeables ; le véritable paramètre d'intérêt ici est plutôt le produit rp. Ceci est une conséquence de la structure en arbre des petites composantes dans l'ensemble CM. Comme nous le verrons au début du chapitre 4, il en est tout autrement lorsque les graphes dérogent d'une telle structure en arbre.

1.3.3 Transition de phase et taille de la composante géante

Étudions plus en détails l'émergence de la composante géante. Pour ce faire, définissons a comme étant la probabilité qu'un lien *ne mène pas* à la composante géante. Cette probabilité s'obtient en demandant qu'elle respecte la condition d'auto-cohérence suivante : si un lien incident sur un nœud quelconque ne mène pas à la composante géante, alors aucun des liens sortants de ce même nœud ne doit non plus mener à la composante géante. Étant donné la



(a) Régime sous-critique $\langle k \rangle_{\text{exp}} = 0.45$

(b) Régime surcritique $\langle k \rangle_{\text{exp}} = 2.0$

FIGURE 1.10 – Solution graphique de l'équation (1.64) pour les régimes sous-critique et surcritique. On y voit d'une part que la fonction $f(x)$ est monotone croissante dans l'intervalle $[0, 1]$, et par conséquent qu'elle ne croise la diagonale x qu'au plus une seule fois dans $[0, 1)$. Ce croisement se produit que lorsque $f'(1) > 1$ et implique la présence d'une composante géante. D'autre part, les diagrammes *cobweb* démontrent qu'en régime sous-critique, l'application (1.66) converge vers le point fixe stable $a = 1$, alors qu'en régime surcritique, une autre solution $a < 1$ devient le point fixe stable de cette application dans $[0, 1)$ et ceci aux dépens du point fixe $a = 1$ qui perd sa stabilité.

structure en arbre des petites composantes (c.-à-d. l'absence de boucle), cette condition peut s'écrire mathématiquement comme

$$a = \sum_{k_s=0}^{\infty} P_s(k_s) a^{k_s} = f(a) \; . \tag{1.64}$$

Autrement dit, si un nœud ne mène pas à la composante géante via ses k_s liens sortants, ce qui se produit avec une probabilité a^{k_s}, alors le lien suivit pour arriver à ce nœud ne mène pas à composante géante. Puisque la probabilité a est définie pour un lien quelconque, nous moyennons la valeur de a^{k_s} avec la distribution des degrés sortants, $\{P_s(k_s)\}_{k_s \in \mathbb{N}}$, ce qui correspond à évaluer la fonction génératrice $f(x)$ au point $x = a$. S'ensuit alors l'équation (1.64).

Par définition, $a = 1$ en l'absence de composante géante, et nous constatons qu'il s'agit bien d'une solution de l'équation (1.64) puisque $f(x)$ génère une distribution de probabilité [c.-à-d. $f(1) = 1$]. Il est toutefois possible que cette équation ait une autre solution dans $[0, 1]$ et, le cas échéant, cette solution est unique, par convexité, dans l'intervalle ouvert $[0, 1)$. Notons d'abord qu'en plus d'être convexe, la fonction $f(x)$ est *monotone croissante* dans $[0, 1]$ puisque ses coefficients représentent des probabilités et sont par conséquent positifs (voir la figure 1.9). Notons également qu'à $x = 0$,

$$0 \leq f(0) = P_s(0) \leq 1 \; .$$

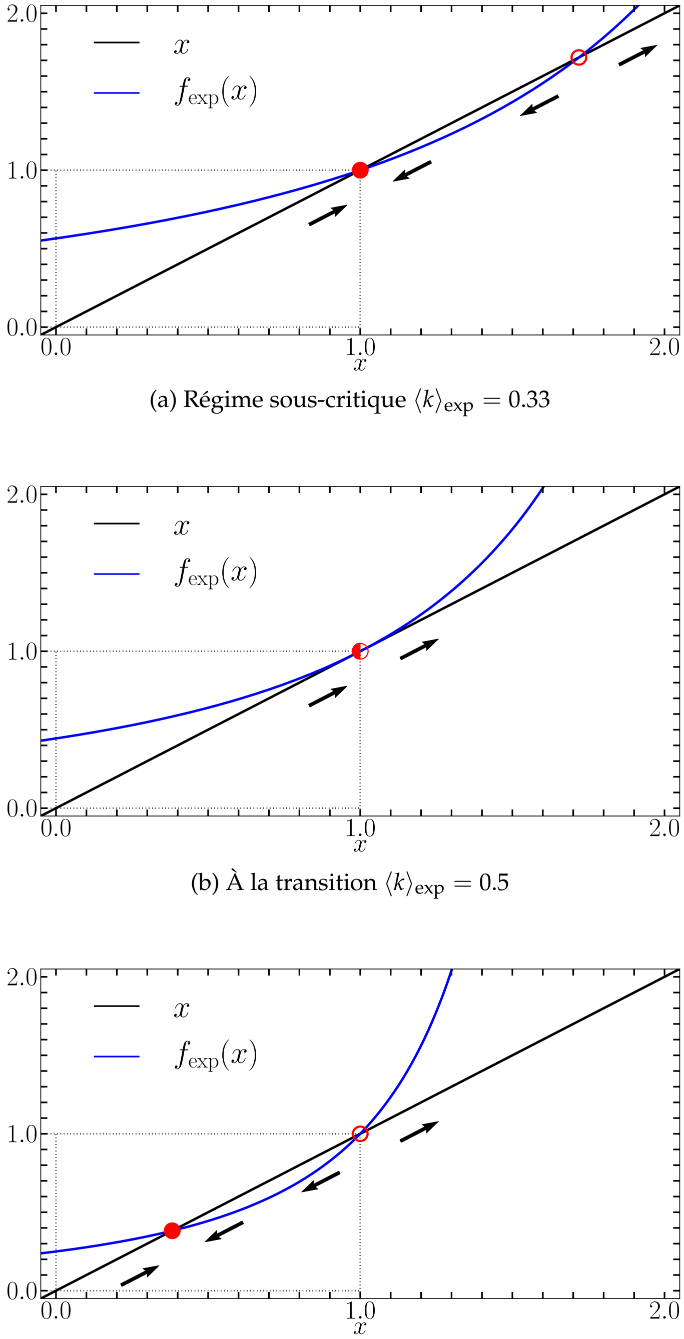
Ainsi, $f(x)$ croît de façon monotone de $f(0) = P_s(0)$ en $x = 0$ jusqu'à $f(1) = 1$ en $x = 1$, ce qui implique que $f(x)$ peut croiser une fois la diagonale $f(x) = x$ dans l'intervalle $[0, 1)$. Un tel croisement correspond à l'existence d'une composante géante puisqu'il existe alors une solution $a < 1$, ce qui implique qu'une fraction $1 - a$ des liens mènent à une composante extensive. En tenant compte de ces propriétés, il est facile de se convaincre qu'un tel croisement n'aura lieu que si la pente de $f(x)$ en $x = 1$ est supérieure à 1 (voir par exemple la figure 1.10 où la diagonale en noir possède une pente en $x = 1$ égale à 1). Ceci correspond au critère pour l'existence d'une composante géante discuté aux sections précédentes, soit

$$f'(1) = \frac{\langle k^2 \rangle - \langle k \rangle}{\langle k \rangle} > 1 \; . \tag{1.65}$$

Les solutions de l'équation (1.64) dans l'intervalle $[0, 1]$ peuvent également être interprétées en termes des points fixes de l'application associée

$$a^{(i+1)} = f\big(a^{(i)}\big) \tag{1.66}$$

avec une condition initiale quelconque $a^{(0)} \in [0, 1)$ et $i \in \mathbb{N}$. En effet, puisque $f(x)$ génère une distribution de probabilité, $a = 1$ est toujours un point fixe de cette application,



(a) Régime sous-critique $\langle k \rangle_{\mathrm{exp}} = 0.33$

(b) À la transition $\langle k \rangle_{\mathrm{exp}} = 0.5$

(c) Régime surcritique $\langle k \rangle_{\mathrm{exp}} = 1.0$

FIGURE 1.11 – Illustration de la bifurcation transcritique correspondant à l'émergence de la composante géante dans l'ensemble CM. Un point fixe est indiqué par un cercle rouge plein (●) s'il est stable, un cercle vide (○) s'il est instable, et un cercle semi-plein (◐) s'il est semi-stable. Les flèches indiquent la direction vers laquelle se déplacera les itérés de l'application (1.66).

et nous démontrons[24] à la figure 1.10a que ce point fixe est globalement stable dans l'intervalle $[0,1]$ dans le régime sous-critique. Lorsqu'il existe une autre solution dans $[0,1]$, c.-à-d. lorsque $f(x)$ croise la diagonale x dans $[0,1)$, nous démontrons à la figure 1.10b que ce « nouveau » point fixe est globalement stable dans $[0,1)$ et ce aux dépens du point fixe $a = 1$ qui perd sa stabilité. En d'autres termes, la transition de phase associée à l'émergence d'une composante extensive dans l'ensemble CM correspond à une bifurcation transcritique de l'application (1.66) dans laquelle la nature de la stabilité du point fixe $a = 1$ change. Par conséquent, une analyse linéaire de la stabilité de ce point fixe [188] nous permet d'obtenir le critère suivant pour l'existence d'une composante extensive

$$f'(1) = \frac{\langle k^2 \rangle - \langle k \rangle}{\langle k \rangle} > 1 \, , \tag{1.67}$$

ce qui est conforme à ce que nous avons obtenu précédemment. La figure 1.11 illustre cette bifurcation transcritique et le changement de stabilité du point fixe $a = 1$. En plus de résumer les propriétés de $f(x)$ (ex. sa croissance monotone, son passage obligé par $f(1) = 1$, etc.) de même que le changement de stabilité du point fixe $a = 1$, cette figure dévoile la nature canonique de la bifurcation transcritique en confirmant qu'il s'agit bien de deux points fixes « s'échangeant » leur stabilité respective en se fusionnant au moment de la transition de phase.

Connaissant maintenant la valeur de la probabilité a [c.-à-d. celle correspondant au point fixe stable de l'application (1.66)], nous sommes en mesure de calculer la probabilité qu'un nœud quelconque ne fasse *pas* partie de la composante géante. Pour ce faire, notons qu'un nœud quelconque dont le degré est égal à k ne fait pas partie de la composante géante avec une probabilité a^k (aucun de ses liens ne mène à celle-ci). À l'opposée, ce nœud fera partie de la composante géante avec la probabilité complémentaire $1 - a^k$. En pondérant cette dernière probabilité par la fraction des nœuds ayant un degré égal à k, nous obtenons la probabilité qu'un nœud choisi au hasard dans un graphe de l'ensemble CM fasse partie de la composante géante

$$\sum_{k=0}^{\infty} P(k) \left[1 - a^k \right] = 1 - g(a) \, . \tag{1.68}$$

Or, puisque ce nœud est choisi aléatoirement, cette probabilité correspond précisément à la fraction des nœuds faisant partie de la composante géante, $\mathcal{S}$, et par conséquent nous avons

$$\mathcal{S} = 1 - g(a) \, . \tag{1.69}$$

La figure 1.12 montre le comportement de $\mathcal{S}$ pour l'ensemble CM dans lequel les degrés sont distribués exponentiellement.

[24] Bien que cette démonstration soit faite en utilisant une distribution des degrés particulière (voir la section 1.3.5), le comportement est suffisamment générique de sorte que les conclusions tirées peuvent être considérées comme générales.

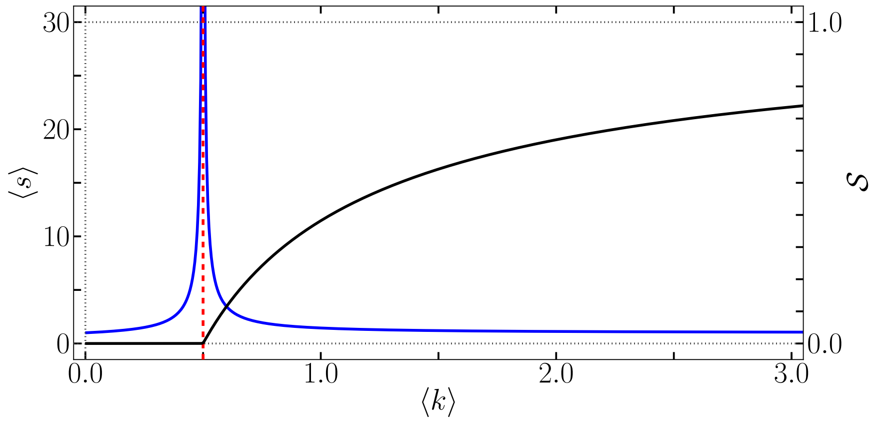


FIGURE 1.12 – Taille moyenne des petites composantes, $\langle s\rangle$ (en bleu), et taille relative de la composante géante, $\mathcal{S}$ (en noir), pour l'ensemble CM avec une distribution des degrés exponentielle. La ligne pointillée rouge montre la valeur du degré moyen correspondant à la transition de phase ($\langle k\rangle = 0.5$). Les courbes bleue et noire ont été obtenues en solutionnant les équations (1.64), (1.69) et (1.84). Voir la section 1.3.5 pour plus de détails sur la distribution des degrés exponentielle.

Retour sur les graphes aléatoires de type Erdős-Rényi

Revenons un instant aux graphes aléatoires de type Erdős-Rényi présentés à la section 1.2. Nous avons vu que dans la limite $N \to \infty$, l'ensemble $\mathcal{G}_{N,p}$ possède une distribution des degrés binomiale qui tend vers une distribution de Poisson lorsque le degré moyen des nœuds, z, est fixe. Des équations (1.42) et (1.50), nous trouvons que pour cet ensemble

$$g_{\text{ER}}(x) = f_{\text{ER}}(x) = \mathrm{e}^{z(x-1)} \; . \tag{1.70}$$

Autrement dit, l'indépendance de l'existence des liens dans l'ensemble $\mathcal{G}_{N,p}$ implique que la distribution des degrés et la distribution des degrés sortants sont identiques. Les équations (1.64) et (1.69) deviennent alors

$$a_{\text{ER}} = \mathrm{e}^{z(a_{\text{ER}}-1)} \tag{1.71}$$

et

$$\mathcal{S}_{\text{ER}} = 1 - \mathrm{e}^{z(a_{\text{ER}}-1)} \; , \tag{1.72}$$

qui peuvent être combinées pour réobtenir l'équation (1.24)

$$\mathcal{S}_{\text{ER}} = 1 - \mathrm{e}^{-z\mathcal{S}_{\text{ER}}} \; . \tag{1.73}$$

Similairement, l'équation (1.65) nous permet de réobtenir la condition pour l'existence d'une composante géante

$$f'_{\text{ER}}(1) = z > 1 \; , \tag{1.74}$$

et la discussion entourant l'équation (1.56) nous permet d'interpréter cette condition comme correspondant à la situation où chaque nouveau nœud rencontré en parcourant le graphe mène à plus d'un nouveau nœud en moyenne. Ainsi, le nombre de nœuds accessibles à partir d'un nœud quelconque croît exponentiellement avec la distance à ce nœud, ce qui correspond à l'existence d'une composante extensive. Ces deux résultats démontrent donc que l'ensemble $\mathcal{G}_{N,p}$ dans la limite appropriée n'est qu'un cas spécial du modèle des configurations.

1.3.4 Taille des petites composantes

L'approche basée sur les fonctions génératrices nous permet également d'obtenir la distribution de la taille des petites composantes dans les graphes de l'ensemble CM. Pour ce faire, définissons les fonctions génératrices

$$A(x) = \sum_{t=1}^{\infty} \rho(t) x^t \tag{1.75}$$

$$K(x) = \sum_{s=1}^{\infty} \eta(s) x^s \ , \tag{1.76}$$

où $\rho(t)$ est la probabilité qu'un lien quelconque mène à une composante contenant t nœuds, et $\eta(s)$ est la probabilité qu'un nœud quelconque fasse partie d'une composante de s nœuds (incluant ce nœud initial). Pour faciliter le calcul, définissons la probabilité intermédiaire $Q_1(t|k)$ qu'un total de t nœuds se trouvent *en aval* d'un nœud dont le degré sortant est k (c.-à-d. un total de t nœuds sont atteignables en suivant les liens sortants de ce nœud). Étant donné la structure en arbre des petites composantes dans la limite $N \to \infty$, nous savons que le nombre de nœuds atteignables via chacun de ces liens sortants est également distribué selon $\{\rho(t)\}_{t\in\mathbb{N}}$. Par conséquent, il y aura t nœuds en aval d'un nœud atteint par un de ses liens si le nombre de nœuds atteignables via ses k liens sortants—menant respectivement vers $t_1', t_2', \ldots, t_k'$ nœuds—est égal à t

$$Q_1(t|k) = \begin{cases} 1 & t = k = 0 \\ \displaystyle\sum_{t_1'=1}^{\infty} \rho(t_1') \ldots \sum_{t_k'=1}^{\infty} \rho(t_k') \delta\left(t, \sum_{m=1}^{k} t_m'\right) & t \geq k > 0 \\ 0 & \text{autrement} \end{cases} ,$$

où nous avons explicité le cas où le degré sortant du nœud atteint via un de ses liens est nul. En pondérant $Q_1(t|k)$ par la probabilité que ce nœud ait un degré sortant égal à k, nous obtenons une expression récursive pour $\rho(t)$

$$\rho(t) = \sum_{k=0}^{\infty} Q_1(t-1|k) P_{\mathrm{s}}(k) \ . \tag{1.77}$$

La fonction génératrice $A(x)$ devient alors

$$\begin{aligned}
A(x) &= \sum_{t=1}^{\infty}\sum_{k=0}^{\infty} Q_1(t-1|k)\tilde{P}(k)x^t \\
&= x\sum_{k=0}^{\infty}\tilde{P}(k)\sum_{t=1}^{\infty}x^{t-1}\left[\sum_{t_1'=1}^{\infty}\rho(t_1')\ldots\sum_{t_k'=1}^{\infty}\rho(t_k')\delta\left(t-1,\sum_{m=1}^{k}t_m'\right)\right] \\
&= x\sum_{k=0}^{\infty}\tilde{P}(k)\left[\sum_{t'=1}^{\infty}\rho(t')x^{t'}\right]^k \\
&= x\sum_{k=0}^{\infty}\tilde{P}(k)\big[A(x)\big]^k \\
&= xf\big(A(x)\big)\ ,
\end{aligned} \tag{1.78}$$

où nous avons utilisé l'équation (1.50). Autrement dit, la fonction génératrice $A(x)$ est le point fixe fonctionnel de l'équation (1.78). En effet, en itérant $n+1$ fois l'équation

$$A^{(i+1)}(x) = xf\big(A^{(i)}(x)\big) \tag{1.79}$$

à partir de la condition initiale $A^{(0)}(x) = 1$ et avec $i \in \mathbb{N}$, nous obtenons exactement les n premiers termes de la distribution $\{\rho(t)\}_{t\in\mathbb{N}}$.

Similairement, définissons la probabilité $Q_0(s|k)$ qu'un nœud mène ultimement à un total de s nœuds via ses k liens. En suivant un même raisonnement que pour $Q_1(t|k)$, nous obtenons

$$Q_0(s|k) = \begin{cases} 1 & s = k = 0 \\ \displaystyle\sum_{t_1'=1}^{\infty}\rho(t_1')\ldots\sum_{t_k'=1}^{\infty}\rho(t_k')\delta\left(s,\sum_{m=1}^{k}t_m'\right) & s \geq k > 0 \\ 0 & \text{autrement} \end{cases} ,$$

à la seule différence qu'ici k est le degré total du nœud et non son degré sortant. La probabilité qu'un nœud quelconque fasse partie d'une composante de s nœuds est alors simplement

$$\eta(s) = \sum_{k=0}^{\infty} Q_0(s-1|k)P(k) \tag{1.80}$$

et l'équation (1.76) devient

$$\begin{aligned}
K(x) &= \sum_{s=1}^{\infty}\sum_{k=0}^{\infty} P(k) Q_0(s-1|k) x^s \\
&= x \sum_{k=0}^{\infty} P(k) \sum_{s=1}^{\infty} x^{s-1} \left[\sum_{t'_1=1}^{\infty} \rho(t'_1) \ldots \sum_{t'_k=1}^{\infty} \rho(t'_k) \delta\left(s, \sum_{m=1}^{k} t'_m\right)\right] \\
&= x \sum_{k=0}^{\infty} P(k) \left[\sum_{t'=1}^{\infty} \rho(t') x^{t'} \right]^k \\
&= x \sum_{k=0}^{\infty} P(k) \big[A(x)\big]^k \\
&= x g\big(A(x)\big) \ , \qquad (1.81)
\end{aligned}$$

où nous avons utilisé l'équation (1.42). Ainsi, ayant obtenu $A(x)$ jusqu'à une taille donnée à l'aide de l'équation (1.79), nous obtenons directement la fonction génératrice associée à la distribution de la taille des petites composantes.

Taille moyenne des petites composantes

Notons que nous pouvons obtenir de l'information sur les petites composantes à partir des équations (1.78) et (1.81) sans avoir à itérer l'équation (1.79). Par exemple, il est possible de calculer la taille moyenne des petites composantes, $\langle s \rangle$, sans préalablement en obtenir la distribution. En régime surcritique, la somme des probabilités qu'un nœud fasse partie d'une composante de taille s, $K(1)$, n'est plus normalisée puisqu'une fraction $\mathcal{S}$ des nœuds fait partie de la composante géante. Autrement dit $K(1) = 1 - \mathcal{S}$. Par conséquent, nous utilisons le ratio $K(s)/K(1)$ pour le calcul général de tout moment de la distribution $\{\eta(s)\}_{s\in\mathbb{N}}$. Les moments d'une distribution pouvant être obtenus via la différentiation de sa fonction génératrice associée, nous obtenons

$$\langle s \rangle = \frac{\sum_{s=1}^{\infty} s\eta(s)}{\sum_{s=1}^{\infty} \eta(s)} = \left[\frac{\partial}{\partial x}\frac{K(x)}{K(1)}\right]_{x=1} = \frac{g(a) + g'(a)A'(1)}{1-\mathcal{S}} \ , \qquad (1.82)$$

où le prime dénote la dérivée d'une fonction par rapport à son argument et où nous avons utilisé le fait que $A(1) = a$. En effet, $A(1)$ est la probabilité qu'un lien quelconque mène à une composante non extensive, ce qui correspond précisément[25] à la définition de a. Une forme explicite pour $A'(1)$ s'obtient en dérivant l'équation (1.78)

$$\begin{aligned}
A'(1) &= \left[\frac{\partial A(x)}{\partial x}\right]_{x=1} \\
&= f(a) + f'(a)A'(1) \ , \qquad (1.83)
\end{aligned}$$

[25] Les équations (1.64) et (1.69) s'obtiennent des équations (1.75) et (1.76) en posant $x = 1$.

ce qui nous permet d'obtenir

$$\langle s \rangle = 1 + \frac{\langle k \rangle a^2}{(1-\mathcal{S})[1-f'(a)]} \tag{1.84}$$

en utilisant la relation entre les fonctions $g(x)$ et $f(x)$ [voir l'équation (1.50)] de même que les équations (1.64) et (1.69).

En régime sous-critique, il n'y a pas de composante géante, par conséquent $1-\mathcal{S} = a = 1$ et l'équation (1.84) devient

$$\langle s \rangle = 1 + \frac{\langle k \rangle}{1-f'(1)} \,. \tag{1.85}$$

Or, en théorie de la percolation, la transition de phase correspondant à l'émergence d'une composante extensive est marquée par la divergence de $\langle s \rangle$ [42], ce qui se produit lorsque

$$f'(1) = \frac{\langle k^2 \rangle - \langle k \rangle}{\langle k \rangle} = 1 \,. \tag{1.86}$$

À nouveau, nous retrouvons le critère marquant la différence entre les régimes sous-critique et surcritique. La figure 1.12 illustre cette divergence.

1.3.5 Distribution des degrés exponentielle

Le comportement des fonctions génératrices introduites dans les sections précédentes a été illustré à plusieurs reprises en utilisant une distribution des degrés exponentielle [26]

$$P_{\text{exp}}(k) = (1-\mathrm{e}^{-\lambda})\mathrm{e}^{-\lambda k} \,, \qquad k \in \mathbb{N},\ \lambda \in \mathbb{R}^+ \,, \tag{1.87}$$

où $\mathbb{N}$ est l'ensemble des nombres naturels (incluant zéro), et $\mathbb{R}^+$ est l'ensemble des nombres réels positifs. La fonction génératrice associée à cette distribution des degrés s'obtient en sommant la série géométrique

$$g_{\text{exp}}(x) = (1-\mathrm{e}^{-\lambda}) \sum_{k=0}^{\infty} \left[x\mathrm{e}^{-\lambda}\right]^k = \frac{1-\mathrm{e}^{-\lambda}}{1-x\mathrm{e}^{-\lambda}} \,, \tag{1.88}$$

où nous avons supposé que $|x\mathrm{e}^{-\lambda}| < 1$, puisque les fonctions génératrices ne nous sont utiles dans dans l'intervalle $x \in [0,1]$. De l'équation (1.46), nous obtenons par différentiation le degré moyen

$$\langle k \rangle_{\text{exp}} = \frac{1}{\mathrm{e}^{\lambda}-1} \,, \tag{1.89}$$

[26] À proprement parler, il ne s'agit pas de la distribution exponentielle qui correspond habituellement à une distribution continue définie pour les nombres réels positifs, mais plutôt de la distribution géométrique [184]. Dans le présent contexte, cette distribution peut-être interprétée de la manière suivante. Soit un nœud ayant un degré initial nul auquel on ajoute 1 à son degré tant et aussi longtemps qu'une épreuve de Bernouilli retourne un « succès » (avec une probabilité $\mathrm{e}^{-\lambda}$). Répétant ceci pour tous les nœuds du graphe (formellement un nombre infini), celui-ci sera caractérisé par la distribution des degrés donnée par l'équation (1.87). Nous avons choisi cette distribution car elle offre un compromis satisfaisant entre réalisme (distribution asymétrique) et simplicité mathématique (un seul paramètre libre facilement interprétable).

ce qui nous permet de fixer le paramètre λ via

$$\lambda = \ln\left(1 + \frac{1}{\langle k\rangle_{\text{exp}}}\right) \ , \tag{1.90}$$

et donc d'utiliser $\langle k\rangle_{\text{exp}}$ comme paramètre de contrôle. De l'équation (1.50), nous obtenons

$$f_{\text{exp}}(x) = \left[\frac{1 - \mathrm{e}^{-\lambda}}{1 - x\mathrm{e}^{-\lambda}}\right]^2 \ , \tag{1.91}$$

ce qui nous permet de conclure [cf. équation (1.40)] qu'il y aura une composante extensive lorsque

$$\langle k\rangle_{\text{exp}} \geq \frac{1}{2} \ . \tag{1.92}$$

Il pourra donc y avoir une composante géante malgré que les nœuds aient en moyenne un degré inférieur à 1. Ce résultat contraste avec la condition $\langle k\rangle \geq 1$ pour les graphes aléatoires de type Erdős-Rényi (section 1.2) et illustre bien l'influence du second moment de la distribution des degrés sur le point marquant la transition de phase [cf. équation (1.40)].

Ce résultat quelque peu surprenant peut également être interprété en rappelant que les nœuds de degré nul n'affectent pas la présence de la composante géante, mais seulement sa taille relative, tel que mentionné à la section 1.3.2. Ainsi, via la transformation donnée à l'équation (1.60), le degré moyen des nœuds dont le degré est non nul est en fait

$$\langle\tilde{k}\rangle \equiv \left[\frac{\partial\tilde{g}(x)}{\partial x}\right]_{x=1} = \frac{g'(1)}{1 - g(0)} = \frac{\langle k\rangle}{1 - g(0)} \ , \tag{1.93}$$

ce qui est nécessairement supérieur ou égal à un. Dans le cas de la distribution exponentielle, le degré moyen des nœuds dont le degré est non nul est $\langle k\rangle_{\text{exp}} + 1$. Lorsqu'elle existe, la composante géante sera donc formée à partir de nœuds dont le degré moyen est supérieur ou égal à 3/2, conformément à l'intuition.

1.3.6 L'ensemble CM en tant que modèle de systèmes réels

À la section 1.2.4, nous avons énuméré les lacunes de l'ensemble $\mathcal{G}_{N,p}$ en tant que modèle pour reproduire les propriétés des graphes tirés de systèmes complexes réels. Parmi celles-ci il y avait la distribution des degrés, et nous avons présenté l'ensemble CM pour remédier à cette situation. Toutefois, comme nous l'avons montré à plusieurs reprises dans cette dernière section, l'ensemble CM conserve plusieurs des lacunes de l'ensemble $\mathcal{G}_{N,p}$ telles que l'absence d'agrégation [cf. équation (1.57)], l'absence de corrélation [cf. équation (1.48)], et l'absence de structure communautaire. Bien que l'inclusion de la distribution des degrés soit un pas considérable vers une modélisation réaliste des systèmes complexes réels, beaucoup de travail reste encore à être accompli. Comme nous l'avons mentionné au début de la section 1.3, le modèle des configurations est très malléable et plusieurs études ont proposé des versions plus générales de l'ensemble CM où certaines propriétés réalistes y sont intégrées [6, 10, 23, 77, 80, 84, 92, 98, 102, 108, 125, 127, 135, 137, 138, 143, 151, 177, 180, 192, 193, 202]. Les travaux faisant l'objet de cette thèse suivent cette tendance, une version très générale de l'ensemble CM a été développée de laquelle la plupart des modèles proposés à ce jour sont des cas spéciaux (voir les chapitres 3 et 4).

Chapitre 2

Théorie I : Solution exacte de la percolation par liens sur graphes arbitraires de petite taille

Article original :

Exact solution of bond percolation on small arbitrary graphs

Antoine Allard [1], Laurent Hébert-Dufresne, Pierre-André Noël,
Vincent Marceau et Louis J. Dubé

Département de Physique, de Génie Physique, et d'Optique,
Université Laval, Québec (Qc), Canada G1V 0A6

Référence : Europhysics Letters, **98** 16001 (2012) [7]

© 2012 EPLA (§2.2 –§2.8) [2]

[1] Le développement théorique, les simulations numériques de même que la rédaction du manuscrit sont essentiellement les fruits du travail du premier auteur. Les autres auteurs ont participé à l'élaboration générale du projet de même qu'à la révision du manuscrit.

[2] Ces sections contiennent le contenu original de l'article. Celui-ci n'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

Cet article présente une version généralisée des équations de récurrence (1.28) et (1.29) dans laquelle les noeuds sont différienciés à l'aide de types, ce qui permet de faire varier la probabilité d'existence des liens en fonction du type des noeuds qu'ils joignent. Lorsqu'à chaque noeud est attribué un type unique, il est possible de prescrire une structure fixe au graphe pour ensuite obtenir la distribution exacte de la taille des composantes si les liens sont retirés aléatoirement. Autrement dit, les équations présentées dans cet article permettent de solutionner exactement la percolation par liens sur de petits graphes dont la structure est fixe et quelconque (c.-à-d. prescrite par une matrice d'adjacence donnée).

Cet article explique également comment les équations (1.28) et (1.29) peuvent être utilisées pour prédire la fragmentation des graphes de l'ensemble $\mathcal{G}_{N,p}$ (c.-à-d. le nombre de composantes et leur taille respective). Cette application fut inspirée par les travaux de Pierre Désesquelles (Université Paris-Sud) qui, lors d'un colloque au département de physique de l'Université Laval, montra comment les prédictions analogues pour l'ensemble $\mathcal{G}_{N,M}$ pouvaient être utilisées comme hypothèse nulle dans l'analyse des résultats d'expériences de fragmentation nucléaire [53, 54]. Il nous avait alors semblé qu'il serait possible de faire de même avec le pendant *canonique* de l'ensemble $\mathcal{G}_{N,M}$ et cet article présente le résultat.

La version multitype des équations de récurrences (1.28) et (1.29) a été développée concomitamment à la généralisation de l'ensemble CM présentée au chapitre 3 et, à l'origine, le tout était destiné à être présenté dans un seul article. Toutefois, l'extension à la percolation par liens sur des graphes arbitraires, de même qu'à la fragmentation de l'ensemble $\mathcal{G}_{N,p}$ nous a incité à y consacrer une publication à part. Cet article est donc un prélude à l'article présenté au chapitre 3, avec en aparté quelques extensions que nous jugeons intéressantes.

Quantity	Description	Definition
M	Number of types of nodes	Sec. 2.4
$\boldsymbol{n}$	Composition of a graph: there are n_i nodes of type i (for every $i = 1, \ldots, M$)	Sec. 2.4
$i \to j$	Directed edge from a node of type i towards a node of type j	Sec. 2.4
p_{ij}	Probability that a $i \to j$ edge exists	Sec. 2.4
$Q_i(\boldsymbol{l}\|\boldsymbol{n})$	Probability that the component reached from a node of type i contains $\boldsymbol{l}$ nodes given that the graph originally contains $\boldsymbol{n}$ nodes	Sec. 2.4
$\mathbf{A}$	Adjacency matrix	Sec. 2.5
T	Probability for an edge to remain after the random removal of edges in an arbitrary graph	Sec. 2.5
q_k	Probability of finding a component of size k, regardless of the identity of the nodes, from a randomly chosen node	Eq. (2.4)
$\boldsymbol{r}$	Partition of a graph in which the i-th component contains r_i nodes	Sec. 2.6
$s_{\boldsymbol{r}}$	Number of ways n nodes can be divided into $\boldsymbol{r}$ components	Sec. 2.6
$\mathcal{P}_n$	Integer partition of n	Sec. 2.6
$P(\boldsymbol{r})$	Probability that a graph is divided into $\boldsymbol{r}$ components	Sec. 2.6

TABLE 2.1 – Glossary of the major mathematical objects defined in this chapter.

2.2 Abstract

We introduce a set of iterative equations that exactly solves the size distribution of components on small arbitrary graphs after the random removal of edges. We also demonstrate how these equations can be used to predict the distribution of the node partitions (i.e., the constrained distribution of the size of each component) in undirected graphs. Besides opening the way to the theoretical prediction of percolation on arbitrary graphs of large but finite size, we show how our results find application in graph theory, epidemiology, percolation and fragmentation theory.

2.3 Introduction

Percolation on graphs is the study of the behavior of components in graphs whose nodes/edges are removed randomly. It has received a lot of attention recently for its various applications in many disciplines. Among them, let us mention the modeling of epidemic propagation [14, 50, 124, 170] where the size distribution of components corresponds to the outbreak size distribution. The same distribution is also related to the size and composition of fragments in nuclear multifragmentation [49, 53, 152, 191]. Finally, the study of the percolation threshold allows an assessment of the robustness (or reliability) of real networks to the failure of their nodes or edges [4, 34, 37, 74, 78, 101]. Alongside the intrinsic theoretical interest, this type of studies has triggered the development of increasingly realistic and complex models (see [44, 58, 144], and references therein).

In the quest for ever more realistic models, a promising idea is to consider real networks not at the level of the nodes, but at higher levels of organization such as motifs, subgraphs or communities. It has recently been proposed that this perspective could help unify and explain many of the universal properties found in real networks [88]. From the modeling perspective, using motifs as the fundamental building blocks of graphs has allowed to relax some of the limiting assumptions behind existing bond percolation models [6, 80, 98, 138]. This has effectively extended the class of models for which exact results can be obtained.

The advantages gained come at a price however: one needs to solve beforehand the distribution of the size of components in these motifs. While this can be done systematically for Erdős-Rényi subgraphs (or cliques) [78, 138], the general problem must be solved by hand on a case-by-case basis. Hence, one is either limited to very small graphs or one has to rely on numerical simulations.

In this Letter, we introduce a set of iterative equations that exactly compute the size distribution of components in a multitype version of Erdős-Rényi graphs. In the case where a different type is assigned to each node of a graph, the equations produce this distribution for any small arbitrary graphs, defined by an asymmetric nonnegative adjacency matrix, after the random removal of its edges. These equations therefore provide a systematic way to compute the required distributions in the motif-based bond percolation models mentioned previously.

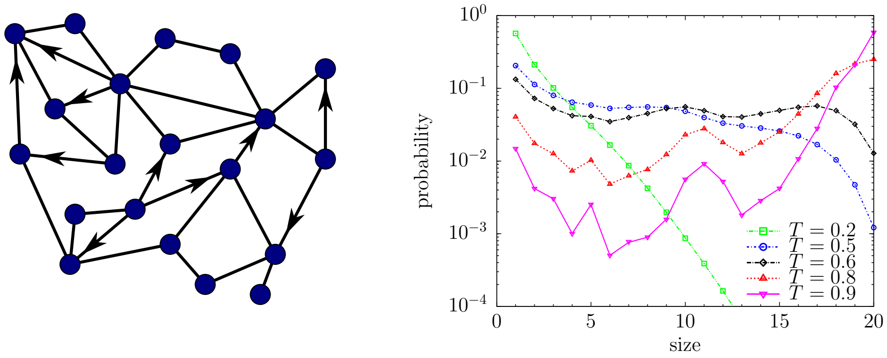
We further show how these equations naturally offer a method to count the number of labeled multitype graphs with a given number of nodes and edges. We also explain how they can be used to study bond percolation on periodic infinite lattices. We finally demonstrate how these equations allow the exact calculation of the constrained distribution of the size of each component, or the partition of nodes, in undirected graphs. It is suggested that this provides a null model for fragmentation processes.

2.4 Percolation on multitype random graphs

Let us consider a multitype generalization of the random graph model $\mathcal{G}_{n,p}$ [78]. These graphs are composed of n nodes and an edge exists between any two nodes with probability p regardless of other potential edges. We generalize this model by labelling nodes using M types such that a graph of size $\boldsymbol{n} \equiv (n_1, \ldots, n_M)^\mathsf{T}$ is composed of n_i type-i nodes (with $i = 1, \ldots, M$). Directed edges from type-i nodes to type-j nodes (noted $i \to j$) exist with probability p_{ij} independently of one another. p_{ij} need not be equal to p_{ji} since edges are directed. Results statistically equivalent to undirected graphs — where an undirected edge exists with a given probability — are obtained in the symmetric case ($p_{ij} = p_{ji}$). While the usefulness of this multitype generalization will become manifest in the next section, multitype Erdős-Rényi graphs can be used as a first approximation of structures in which correlations exist in the way nodes are connected (with an appropriate choice of p_{ij}).

We define $Q_i(\boldsymbol{l}|\boldsymbol{n})$ as the probability that the component reached from a type-i node contains $\boldsymbol{l} \equiv (l_1, \ldots, l_M)^\mathsf{T}$ nodes (including the initial type-i node) given that the graph contains $\boldsymbol{n}$ nodes. Because of the presence of directed edges, we extend the definition of a component to all nodes *accessible* from a given node. Thus, node A being in the component reached from node B does not imply that the converse is true. In the same spirit of [78, 138], we now derive two recurrence equations allowing for the explicit calculation of $Q_i(\boldsymbol{l}|\boldsymbol{n})$.

Let us consider a graph of size $\boldsymbol{n}$, a component of size $\boldsymbol{l}$, and an initial node of type i. We note δ_{ij} the Kronecker delta. Among the $n_j - \delta_{ij}$ nodes of type j (excluding the initial node of type i), there are $\binom{n_j-\delta_{ij}}{l_j-\delta_{ij}}$ ways to choose the l_j type-j nodes that are part of the component. These l_j type-j nodes will not lead to any of the remaining $\boldsymbol{n} - \boldsymbol{l}$ nodes that are not part of the component with probability $\prod_k q_{jk}^{l_j(n_k-l_k)}$, where $q_{jk} \equiv 1 - p_{jk}$. In other words, this is the probability that no directed edge $j \to k$ exists between the l_j type-j nodes in the component and the $\boldsymbol{n} - \boldsymbol{l}$ nodes of every other types excluded from it. Repeating this procedure for the other types of nodes in the component, together with the observation that the $\boldsymbol{l}$ nodes form



(a)

(b)

FIGURE 2.1 – (a) Example of an arbitrary graph of 20 nodes. Although not shown here, graphs with multiple edges could as easily be considered by our method. (b) Distribution of the size of the component reached from a randomly chosen node in the graph shown in (a) where edges have been removed with probability $1 - T$. Lines were obtained using eqs. (2.2)–(2.4), and symbols were obtained by performing 10^7 simulations. The distributions are discrete; lines have been added to guide the eye.

a component with probability $Q_i(\boldsymbol{l}|\boldsymbol{l})$, we have that

$$Q_i(\boldsymbol{l}|\boldsymbol{n}) = Q_i(\boldsymbol{l}|\boldsymbol{l}) \prod_{jk} \binom{n_j - \delta_{ij}}{l_j - \delta_{ij}} q_{jk}^{l_j(n_k - l_k)} \; . \tag{2.1}$$

That is, by knowing the probability of finding a component of size $\boldsymbol{l}$ in a graph of size $\boldsymbol{l}$, eq. (2.1) computes the probability of finding a component of size $\boldsymbol{l}$ in a graph of size $\boldsymbol{n}$ (with $n_j \geq l_j$ for all j). Finally, to obtain $Q_i(\boldsymbol{l}|\boldsymbol{l})$, we note that the distribution $\{Q_i(\boldsymbol{m}|\boldsymbol{l})\}$ must be normalized for a given graph size $\boldsymbol{l}$, hence

$$Q_i(\boldsymbol{l}|\boldsymbol{l}) = 1 - \sum_{\boldsymbol{m}<\boldsymbol{l}} Q_i(\boldsymbol{m}|\boldsymbol{l}) \; . \tag{2.2}$$

The sum covers every possible instances of $\boldsymbol{m}$ such that $m_j \leq l_j$ for all j but excludes the case where $m_j = l_j$ for each j. Starting with the initial condition $Q_i(\boldsymbol{\delta_i}|\boldsymbol{\delta_i}) = 1$, where $\boldsymbol{\delta_i} \equiv (\delta_{i1}, \ldots, \delta_{iM})^{\mathsf{T}}$, we can therefore calculate every coefficient $Q_i(\boldsymbol{l}|\boldsymbol{n})$ using eqs. (2.1)–(2.2) iteratively. In other words, from a graph made out of a single node, eqs. (2.1)–(2.2) extend the graph to the desired size, and keep track of the component size distribution along the way to build the final distribution $Q_i(\boldsymbol{l}|\boldsymbol{n})$. A simple example of such a calculation is given in the Appendix. Setting $M = 1$, we retrieve the recurrence equations presented in [78, 138]. Also, a similar approach has been used to analyse the reliablity of communication networks [101].

Using the identity $1 = \prod_{j,k}(p_{jk} + q_{jk})^{n_j(n_k - \delta_{jk})}$ in eq. (2.2), where $n_j(n_k - \delta_{jk})$ is the maximal number of $j \to k$ edges in the graph, the iteration of eqs. (2.1)–(2.2) yields polynomials whose coefficients have a direct combinatorial interpretation. Indeed, the coefficient in front of $\prod_{j,k} p_{jk}^{a_{jk}} q_{jk}^{b_{jk}}$ in $Q_i(\boldsymbol{l}|\boldsymbol{n})$ is simply the number of distinct ways to reach a component of size $\boldsymbol{l}$ in a graph of size $\boldsymbol{n}$ from a type-i node using a_{jk} existing and b_{jk} absent $j \to k$ edges, respectively. The sum $a_{jk} + b_{jk}$ may be different than $n_j(n_k - \delta_{jk})$ as the existence of some edges may be irrelevant to the component. Of particuliar interest is the symmetric case where $p_{jk} = p_{kj}$ for all j and k for which $Q_i(\boldsymbol{n}|\boldsymbol{n})$ is independent of i and whose coefficients are the number of connected labeled graphs of size $\boldsymbol{n}$. Hence, we see that eqs. (2.1)–(2.2) offer an alternative method to enumerate the number of graphs with a given number of edges and labeled nodes of different types.

2.5 Percolation on arbitrary graphs

By considering that each node of an arbitrary graph belongs to its own type, eqs. (2.1)–(2.2) can exactly predict the outcome of a bond percolation process that has occurred on it. Predicting the outcome here is as precise as knowing the identity of the nodes that have been reached and of the ones that have not. To illustrate this point, let us consider the simplest case where an edge is to be kept with the same probability T during the percolation process[3]. The probability p_{jk} for the edge $j \to k$ to exist then becomes $p_{jk} = 1 - (1-T)^{A_{jk}}$, where $A_{jk} \in \mathbb{N}$ is an element of the adjacency matrix $\mathbf{A}$ corresponding to the number of directed edges from node j to node k. Arbitrary graphs with directed edges and multiple edges can thus be considered with our method.

Because each node belongs to its own type, the elements of the vectors $\boldsymbol{l}$, $\boldsymbol{m}$ and $\boldsymbol{n}$ indicate whether each node is present (value 1) or not (value 0). This allows us to write eq. (2.1) in a simpler and more compact form

$$Q_i(\boldsymbol{l}|\boldsymbol{n}) = Q_i(\boldsymbol{l}|\boldsymbol{l})(1-T)^{\boldsymbol{l}^{\mathsf{T}}\mathbf{A}\bar{\boldsymbol{l}}} , \tag{2.3}$$

where the elements of $\bar{\boldsymbol{l}}$ are defined such that $l_j + \bar{l}_j = n_j$ for all j. That is, $\boldsymbol{l}^{\mathsf{T}}\mathbf{A}\bar{\boldsymbol{l}}$ is the number of outgoing edges that must not exist for the component of size $\boldsymbol{l}$ to be isolated from the rest of the graph. Using eqs. (2.2)–(2.3) together with the initial condition $Q_i(\boldsymbol{\delta_i}|\boldsymbol{\delta_i}) = 1$, we can calculate the exact probability of *each individual outcome* of a percolation process on an arbitrary graph defined by its adjacency matrix $\mathbf{A}$.

To support this claim, fig. 2.1b compares the predictions of eqs. (2.2)–(2.3) with the results of numerical simulations of bond percolation on the graph shown in fig. 2.1a. To lighten the presentation of the results, fig. 2.1b shows the probability q_k of finding a component of size k — regardless of the identity of the nodes — from a randomly chosen node. This quantity

[3] The generality of eqs. (2.1)–(2.2) naturally allows for the use of various cases of type-dependent probabilities of existence of edges, all the way to the most general case where one specific probability is assigned for each direction of each edge.

is computed using

$$q_k = \frac{1}{M} \sum_i \sum_{\boldsymbol{l}} Q_i(\boldsymbol{l}|\boldsymbol{n}) \delta \left(\sum_j l_j - k \right) \tag{2.4}$$

where $\delta(\cdots)$ is the Kronecker delta. We observe an excellent agreement between our theoretical predictions and the numerical results. Although for such a small network no precise percolation threshold can be defined, it is however clear that a qualitative change toward a "giant component" is initiated for $T \sim 0.5$–0.6. Also, the irregular shape of the distribution for some values of T highlights how $Q_i(\boldsymbol{l}|\boldsymbol{n})$ can depend on the precise structure of the graph. This advocates for the importance of developing methods that consider explicitly the structure of the graphs (i.e., the adjacency matrix).

Since eqs. (2.1)–(2.3) consider every possible outcomes of the percolation process, their predictions are exact. However, the calculational burden (e.g., required memory, number of operations) increases very quickly with the number of node types M. In the case of arbitrary graphs, it grows exponentially with the number of nodes. Thus, although eqs. (2.1)–(2.3) are in principle valid for graphs of any size, their use becomes cumbersome for large graphs. With our present computer facilities, a straightforward implementation of eqs. (2.2)–(2.3) have been able to handle graphs of size of the order of 25. A wiser implementation could certainly push this limit somewhat further. When dealing with a given graph, specific features of its structure may however be used to reduce the numerical effort. For instance, the distributions for different modules could be solved separately, and then recombined to obtain $Q_i(\boldsymbol{l}|\boldsymbol{n})$ for the whole graph. Quantum computation of the sort described in [73] may also be a solution for larger graphs.

Despite these limitations, our method compares favorably with an *exact enumeration* method where a computer program explicitly considers each possible edge configuration, and then computes the component size distribution from them. Firstly, the computational demands of this approach scales exponentially with the number of edges L, whereas our approach scales exponentially with the number of nodes. The performance of our method should therefore be comparable to direct enumeration for sparse graphs, and should rapidly surpass it for denser graphs. Secondly, our method yields analytical solutions (i.e., polynomial in p_{jk}) valid for any value of p_{jk}.

Equations (2.2)–(2.3) can also be used to compute the bond percolation threshold of infinite periodic lattices. By virtue of the *triangle-triangle* transformation [172], the percolation threshold is the root of a polynomial related to the connectivity of the basic cell of the lattice, which is a combination of the coefficients of $Q_i(\boldsymbol{l}|\boldsymbol{n})$. Thus our approach offers a systematic and exact way to compute this polynomial for complicated basic cells. The Appendix provides an example. Furthermore, our approach offers a systematic way to obtain the renormalisation-group transformation to estimate the scaling exponents and the bond percolation threshold of infinite lattices (see for instance eq. (3.4) in [165]).

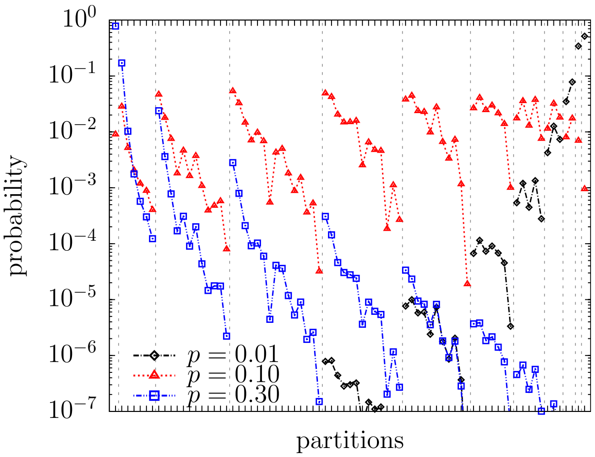


FIGURE 2.2 – Probability of finding each node partition in $\mathcal{G}_{n,p}$ with $n = 12$ and for various values of p. Lines were obtained using eq. (2.7), and symbols were obtained by performing over 5×10^8 simulations. The $|\mathcal{P}_{12}| = 77$ integer partitions of 12 are displayed in an increasing order of the number of components, i.e., from the node partition where there is only one component (noted $\{12\}$) on the left to the case where there are 12 components of one single node (noted $\{1,1,1,1,1,1,1,1,1,1,1,1\}$) on the right. Linebreaks and vertical grey lines indicate where the number of components changes. Node partitions with the same number of components are displayed in an decreasing order of the largest components sizes (e.g., $\{8,2,2\}$ before $\{7,4,1\}$ before $\{7,3,2\}$).

Finally, eqs. (2.1) and (2.3) can be combined to compute $Q_i(\boldsymbol{l}|\boldsymbol{n})$ for graphs where nodes of different types interact through an arbitrary configuration of edges (see [6] for an explicit example). This allows to generate a wide range of realistic subgraphs (or motifs) found for instance in social networks, and to include them in motif-based bond percolation models [6, 98]. As the component size distribution is closely related to the outbreak size distribution, our approach allows to study the spread of infectious diseases in more realistic urban settings [126].

2.6 Predicting node partition distribution

We can also use eqs. (2.1)–(2.3) to calculate the distribution of the number of components (and their size) found in an *undirected* graph after the removal of a fraction of its edges. We restrict ourselves to undirected graphs because only in undirected graphs are components uniquely defined. That is, two nodes will be found in the same component with one unique probability regardless of the starting node. We illustrate how to perform the calculation using the $\mathcal{G}_{n,p}$ model. It should nevertheless be clear that equations for undirected multitype random graphs ($p_{ij} = p_{ji}$ for all i and j) and undirected arbitrary graphs ($\mathbf{A} = \mathbf{A}^\mathsf{T}$) can be derived in a similar manner.

Let us calculate the probability for a random graph composed of n nodes to be split into k components of size $\boldsymbol{r} \equiv (r_1, r_2, \ldots, r_k)$ with $\sum_i r_i = n$. Each component l will be connected, and therefore be a component, with probability $Q(r_l|r_l)$. We have dropped the subscript in $Q(r_l|r_l)$ because the probability to find one *single* connected component in an undirected graph is the same regardless of the starting node. These components will be isolated from one another if none of the possible edges between nodes of different components exist. For the whole graph, this happens with probability $\prod_{i<j}(1-p)^{r_i r_j}$ where p is the probability for an edge to exist between any two nodes.

The remaining step is to count the number of ways the n nodes can be divided into $\boldsymbol{r}$ components, which we note $s_{\boldsymbol{r}}$. This number is in fact equal to the number of ways to put n labeled objects into k nonempty and *unlabeled* containers of size r_1, r_2, ..., and r_k. First, let us point out that the number of ways to put n labeled objects into k *labeled* containers whose sizes are given by $\boldsymbol{r}$ is simply the multinomial coefficient $n!/\prod_i r_i!$. To obtain $s_{\boldsymbol{r}}$, we just remove the redundant configurations due to containers of the same size. Hence, noting $d_m \equiv \sum_i \delta(m - r_i)$ the number of containers of size m, we get

$$s_{\boldsymbol{r}} = \frac{n!}{\left(\prod_m d_m!\right)\left(\prod_i r_i!\right)} \ . \tag{2.5}$$

Note that $s_{\boldsymbol{r}}$ is related to the *Stirling number of the second kind* $\left\{ {n \atop k} \right\}$ [1] giving the number of ways to put n labeled objects into k nonempty and *unlabeled* containers. Indeed $\left\{ {n \atop k} \right\}$ is simply the sum of every $s_{\boldsymbol{r}}$ such that $\boldsymbol{r}$ has k elements

$$\left\{ {n \atop k} \right\} = \sum_{\boldsymbol{r} \in \mathcal{P}_n} s_{\boldsymbol{r}} \delta\big(\dim(\boldsymbol{r}) - k\big) \ , \tag{2.6}$$

where $\mathcal{P}_n$ is the set of *integer partition* of n, i.e., the set of decompositions of n into a unordered sum of integers [85].

Combining these three contributions, we obtain the probability for a random graph composed of n nodes to be split into k components of size $\boldsymbol{r}$ to be

$$P(\boldsymbol{r}) = s_{\boldsymbol{r}} \prod_l Q(r_l|r_l) \prod_{i<j} (1-p)^{r_i r_j} \ . \tag{2.7}$$

To validate eq. (2.7), fig. 2.2 compares its predictions with the results obtained from numerical simulations of the $\mathcal{G}_{n,p}$ model for $n = 12$ and for various values of p. Again, an excellent agreement between our theoretical predictions and the results of the numerical simulations is observed. Figure 2.2 highlights the emergence of a "giant" component — which occurs when $(n-1)p = 1$ in the limit $n \to \infty$ — as the distribution migrates toward partitions with fewer components with increasing p. It also shows that, for a same number of components, the partitions with larger components are more likely to occur in general. Although partly shown in fig. 2.2, this trend holds for all values of $p \in [0, 1]$, and is due to the simple fact that there are more ways (i.e., possible configuration of edges) to build large connected components than small ones.

Our method can also serve as a null model for fragmentation processes. In fact, $P(\boldsymbol{r})$ is the probability for n elements to be distributed among $\boldsymbol{r}$ fragments when bonds occur (or resist) randomly with probability p. Physical correlations can therefore be highlighted by comparing $P(\boldsymbol{r})$ with experimental data. Similar results for $\mathcal{G}_{n,m}$ [62], where the number m of edges (or energy) is fixed rather than its average value $\binom{n}{2}p$, have recently been used in the context of nuclear multifragmentation [53]. As $\mathcal{G}_{n,p}$ and $\mathcal{G}_{n,m}$ can be seen as the "canonical" and "micro-canonical" version of Erdős-Rényi random graphs, the results for $\mathcal{G}_{n,m}$ in [53] can be reobtained using our equations (see the number of connected graphs at the end of the second section). Our method therefore emcompasses previous results, and fills the gap in contexts where the canonical approach is more relevant.

2.7 Conclusion

We have introduced a set of iterative equations that computes the distribution of the size of the components in small random or arbitrary graphs. As directed and multiple edges can naturally be accounted for in the equations, our method is suitable for a wide range of arbitrary graphs. Because the equations consider systematically all possible outcomes of the bond percolation process, their predictions are exact. We have also demonstrated that they can be used to calculate the constrained distribution of the size of each component (i.e., node partition) for undirected small graphs. We have illustrated how these results find applications in various disciplines like graph theory, percolation theory, epidemiology and fragmentation theory. We believe that, despite the increasing unwieldiness of the calculation with the number of nodes, our results open the way to the theoretical prediction of bond percolation on large, but finite, arbitrary graphs.

2.8 Appendix

We perform an explicit calculation of $Q_i(\boldsymbol{l}|\boldsymbol{n})$ to clarify the use of eqs. (2.1)–(2.2) and to illustrate some of our claims.

Let us consider a simple graph composed of 3 nodes of type 0, 1 and 2. The directed $i \to j$ edge exists with probability p_{ij} and there are 6 possible directed edges. We note $q_{ij} = 1 - p_{ij}$. We take the node of type 0 as the starting node without loss of generality as the two other distributions can be obtained by permutation.

The calculation begins with the initial condition $Q_0(1,0,0|1,0,0) = 1$ stating the obvious fact that the probability of finding a component $\boldsymbol{l} = (1,0,0)$ in a graph of size $\boldsymbol{n} = (1,0,0)$ is 1. Equation (2.1) provides the probability of finding the same component but in graphs respectively of size $(1,1,0)$, $(1,0,1)$ and $(1,1,1)$

$$\begin{aligned}
Q_0(1,0,0|1,1,0) &= q_{01} \\
Q_0(1,0,0|1,0,1) &= q_{02} \\
Q_0(1,0,0|1,1,1) &= q_{01}q_{02} \,.
\end{aligned}$$

Using eq. (2.2), we then compute the probability to find a component of size 2 in the graphs of size $(1,1,0)$, $(1,0,1)$

$$\begin{aligned}
Q_0(1,1,0|1,1,0) &= 1 - Q_0(1,0,0|1,1,0) = p_{01} \\
Q_0(1,0,1|1,0,1) &= 1 - Q_0(1,0,0|1,0,1) = p_{02} \,.
\end{aligned}$$

We use once more eq. (2.1) to compute the probability of finding the same components but in a graph of size $\boldsymbol{n} = (1,1,1)$

$$\begin{aligned}
Q_0(1,1,0|1,1,1) &= p_{01}q_{12}q_{02} \\
Q_0(1,0,1|1,1,1) &= p_{02}q_{21}q_{01} \,.
\end{aligned}$$

Finally, the probability of reaching the whole graph of size $(1,1,1)$ is obtained with eq. (2.2)

$$\begin{aligned}
Q_0(1,1,1|1,1,1) &= 1 - q_{01}q_{02} - p_{01}q_{12}q_{02} - p_{02}q_{21}q_{01} \\
&= p_{01}p_{02} + p_{01}p_{12}q_{02} + p_{02}p_{21}q_{01} \,,
\end{aligned}$$

where we have used the identity

$$1 = (p_{01} + q_{01})(p_{02} + q_{02})(p_{12} + q_{12})(p_{21} + q_{21})$$

to obtain a polynomial with positive coefficients. As claimed previously, each term in this last polynomial can be interpreted as a path leading to the component $\boldsymbol{l} = (1,1,1)$, and its coefficient as the number of distinct realisations of such a path.

With these results, we show how to compute the percolation threshold p_c for the infinite triangular lattice using the *triangle-triangle* transformation [172]. Setting $p_{ij} = p$ for all i and j in $Q_0(1,1,1|1,1,1)$ with $q = 1-p$, we retrieve the probability for the three nodes to be connected in some way, that is $p^3 + 3p^2q$. Remember that for undirected graphs, the probability of reaching the entire graph is independent of the starting node. The *triangle-triangle* transformation then stipulates that p_c is the lowest value in [0,1] for which this last probability is equal to the probability q^3 that none of the nodes are connected. Thus, p_c satisfies

$$p_c^3 - 3p_c + 1 = 0 \,,$$

whose only solution in [0,1] is $p_c = 2\sin(\pi/18)$, which is the exact value of the bond percolation threshold for the triangular lattice [190].

This simple example could have been solved without using eqs. (2.1)–(2.2). However, repeating this exercise for graphs of 4, 5 or 6 nodes should convince the reader that a systematic procedure, as provided by eqs. (2.1)–(2.2), quickly becomes necessary.

Chapitre 3

Théorie II : Percolation par liens sur graphes aléatoires

Article original :

Bond percolation on a class of correlated and clustered random graphs

Antoine Allard[1], Laurent Hébert-Dufresne, Pierre-André Noël,
Vincent Marceau et Louis J. Dubé

Département de Physique, de Génie Physique, et d'Optique,
Université Laval, Québec (Qc), Canada G1V 0A6

Référence : J. Phys. A, **45** 405005 (2012) [6]

© 2012 IOP Publishing Ltd (§3.2 –§3.10)[2]

[1] Le développement théorique, les simulations numériques de même que la rédaction du manuscrit sont essentiellement les fruits du travail du premier auteur. Les autres auteurs ont participé à l'élaboration générale du projet de même qu'à la révision du manuscrit.

[2] Ces sections contiennent le contenu original de l'article. Celui-ci n'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

Cet article présente une version multitype de l’ensemble CM dans laquelle les nœuds sont liés via une appartenance commune à un groupe (les groupes étant également distingués à l’aide de types). Au moment de sa publication, il s’agissait d’une des versions les plus générales de l’ensemble CM, voire la plus générale, et nous y avons consacré une section pour en faire la démonstration (voir la section 3.7). En effet, l’approche multitype permettait de reproduire un large éventail de corrélations dans la façon dont les nœuds sont liés, en plus de permettre l’inclusion de l’effet d’agrégation par le biais d’une grande variété de motifs, ceci étant rendu possible grâce aux équations de récurrence introduites au chapitre précédent.

Cette généralisation du CM est une suite logique aux travaux de maîtrise de l’auteur [5, 10] et, à l’origine, ne devait être qu’un court projet en début de doctorat pour « se refaire la main » après une année d’enseignement à l’étranger. Comme nous le montrons aux chapitres suivants, ce projet ouvrit plutôt la porte à plusieurs études intéressantes qui démontrèrent l’utilité et la polyvalence de cette approche théorique.

Quantity	Description	Definition
M	Number of types of nodes	Sec. 3.4.1
Λ	Number of types of groups	Sec. 3.4.1
T	Probability for an infectious node to eventually transmit the disease to one of its susceptible neighbours	Sec. 3.6
p_{rs}	Probability that a directed edge exists from a type-r node to a type-s node in a multitype random motif	Sec. 3.9
w_i	Fraction of nodes of type i	Sec. 3.4.1
$P_i(\boldsymbol{k})$	Probability that a node of type i is connected to k_μ groups of type μ (for each $\mu = 1, \ldots, \Lambda$)	Sec. 3.4.1
$R_\nu(\boldsymbol{n})$	Probability that a group of type ν contains n_j nodes of type j (for each $j = 1, \ldots, M$)	Sec. 3.4.1
$Q_{i\nu}(\boldsymbol{l}\|\boldsymbol{n})$	Probability that $\boldsymbol{l}$ nodes (i.e., l_j type-j nodes, for all j) will eventually be reached from an initial type-i node by following existing edges in a type-ν motif of size $\boldsymbol{n}$	Sec. 3.4.2
$\nu \to i$	Edge stemming from a group of type ν and leading to a node of type j	Sec. 3.5
$\langle b \rangle_B$	Average value of the quantity b over the distribution $B(b)$	Eq. (3.1)
$\xi_{\nu j}(s)$	Average number of $\nu \to j$ edges at a distance s from any node	Eq. (3.5)
$\mathbf{B}$	Matrix whose elements are the average numbers of $\nu \to j$ edges emerging from a type-i node that has been reached via a $\mu \to i$ edge	Eq. (3.7)
$a_{\mu i}$	Probability that a $\mu \to i$ edge does not lead to the giant component	Eq. (3.8)
$\mathcal{P}$	Probability that a randomly chosen node leads to the giant component	Eq. (3.9)
$\bar{a}_{\mu i}$	Probability that a neighbour of a type-i node in a type-μ group is not part of the giant component	Eq. (3.10)
$\mathcal{S}$	Relative size of the giant component	Eq. (3.12)
$\langle s_i \rangle$	Average number of nodes of type i found in small components	Eq. (3.15)
$\theta_{i\nu}(\boldsymbol{x})$	pgf generating the distribution $\{Q_{i\nu}(\boldsymbol{l}\|\boldsymbol{n})\}$ of the outcome of bond percolation, from an initial type-i node, on the motifs corresponding to type-ν groups	Eq. (3.2)
$g_i(\boldsymbol{x})$	pgf generating the distribution of the number of $\nu \to j$ edges emerging from a type-i node	Eq. (3.3)
$f_{\mu i}(\boldsymbol{x})$	pgf generating the distribution of the number of $\nu \to j$ edges emerging from a type-i node which has itself been reached via a $\mu \to i$ edge	Eq. (3.4)
$A_{\mu i}(\boldsymbol{x})$	pgf generating the distribution of the number of edges of each type (i.e., $\nu \to j$ for all ν and j) that are *ahead* of a $\mu \to i$ edge in small components	Eq. (3.13)
$K(\boldsymbol{z})$	pgf generating the distribution of the composition of a small component that is reached from a randomly chosen node	Eq. (3.14)

TABLE 3.1 – Glossary of the major mathematical objects defined in this chapter.

3.2 Abstract

We introduce a formalism for computing bond percolation properties of a class of correlated and clustered random graphs. This class of graphs is a generalization of the Configuration Model where nodes of different types are connected via different types of *hyperedges*, edges that can link more than 2 nodes. We argue that the multitype approach coupled with the use of clustered hyperedges can reproduce a wide spectrum of complex patterns, and thus enhances our capability to model real complex networks. As an illustration of this claim, we use our formalism to highlight unusual behaviors of the size and composition of the components (small and giant) in a synthetic, albeit realistic, social network.

3.3 Introduction

Bond percolation is the study of the size distribution of components in graphs whose edges exist with a given probability. For its theoretical appeal and its varied applications in many contexts, mathematical modelling of bond percolation on random graphs has recently received substantial attention (see [58, 144], and references therein). Within the Configuration Model (CM) paradigm [129, 130], many exact results can be obtained using probability generating functions (PGF) [147]. This analytic tractability however comes at the price of simplifying assumptions on the structure of the graphs.

We introduce a generalization of the CM that encompasses many of the previous improvements published to this day [10, 77, 80, 98, 108, 125, 127, 136, 135, 137, 138, 143, 147, 177, 180, 202], and brings this class of models closer to the behavior of real complex networks. By combining the multitype approach of [10], the analytical method of [7] and the one-mode projection of [138], we argue that our model is able to reproduce a wide range of complex patterns found in real networks.

On the one hand, the multitype approach allows to explicitly prescribe how nodes are connected to one another in a very detailed fashion. By assigning types to nodes – in other words by knowing who is who, and therefore who is connected to whom – several mixing patterns (e.g., assortativity, degree correlation, node segregation), as well as heterogeneous bond occupation probabilities (e.g., partial and/or uneven directionality of edges) can be reproduced. On the other hand, the use of the one-mode projection, coupled with the multitype approach, allows the inclusion of clustering through a myriad of nontrivial motifs, i.e. recurrent, significant patterns of interconnections [128].

This paper is organized as follows. In section 3.4, we introduce the generalization of the CM that explicitly includes various correlations and clustering. We then develop the analytical framework to obtain the bond percolation properties of this graph ensemble in section 3.5. In section 3.6, we validate our formalism — and also illustrate the versatility of our approach —

by comparing its predictions with simulation results on a synthetic, but realistic social network. In section 3.7, we show that many percolation models published in the litterature are special cases of our model. We also highlight how our approach can be useful for studying interdependent or coupled networks [34, 95, 108, 117], and for studying the *weak and strong clustering regimes* [175, 176]. We conclude in section 3.8 and present in 2 Appendices some relevant aspects of the analysis and simulations. Section 3.9 details how a recent method, to analytically compute the distribution of the composition of components for any small arbitrary graphs [7], can be used in our formalism. Section 3.10 gives further details on the numerical simulations.

3.4 Correlated and clustered graph ensemble

We introduce a general class of correlated and clustered random graphs. To preserve the analytical tractability of the CM, we first consider *unclustered multitype bipartite graphs* that are locally tree-like in the large system size limit. Clustering is then incorporated through a projection, analogous to the one-mode projection of [138].

3.4.1 Unclustered multitype bipartite graph ensemble

We call unclustered multitype bipartite graphs a multitype generalization of the bipartite CM [147]. These graphs are composed of M types of "regular nodes" and Λ types of "group nodes" (hereafter referred to as nodes and groups, respectively). Edges only exist between regular nodes and group nodes. In these graphs, a fraction w_i of nodes are of type i, and any given type-i node is connected to k_μ type-μ group nodes (for each $\mu = 1, \ldots, \Lambda$) with a probability $P_i(k_1, \ldots, k_\Lambda) \equiv P_i(\boldsymbol{k})$. Likewise, a randomly chosen type-ν group is connected to n_j type-j nodes (for each $j = 1, \ldots, M$) with a probability $R_\nu(n_1, \ldots, n_M) \equiv R_\nu(\boldsymbol{n})$. In other words, $R_\nu(\boldsymbol{n})$ is the distribution of the composition of type-ν groups. Figure 3.1a gives an example of such graphs. To lighten the notation, it should now be understood that any free latin (resp. greek) index can take any values in $\{1, \ldots, M\}$ (resp. $\{1, \ldots, \Lambda\}$), except if otherwise mentionned.

In the large system size limit, w_i, $P_i(\boldsymbol{k})$ and $R_\nu(\boldsymbol{n})$ fully define a graph ensemble which is totally random in all other respects (stubs are matched randomly). All finite components therefore have a tree-like structure in this limit (the probability of a closed path goes as the inverse of the size of the graph). These quantities are however not independent. To guarantee the consistency of the graph ensemble, they must, for all applicable combinations of i, j and ν, satisfy

$$\frac{w_i \langle k_\nu \rangle_{P_i}}{\langle n_i \rangle_{R_\nu}} = \frac{w_j \langle k_\nu \rangle_{P_j}}{\langle n_j \rangle_{R_\nu}} , \tag{3.1}$$

where $\langle a \rangle_B$ denotes the mean value of a with respect to the distribution B. Simply stated, equation 3.1 asks w_i, $P_i(\boldsymbol{k})$ and $R_\nu(\boldsymbol{n})$ to be chosen such that each node type "forces" the

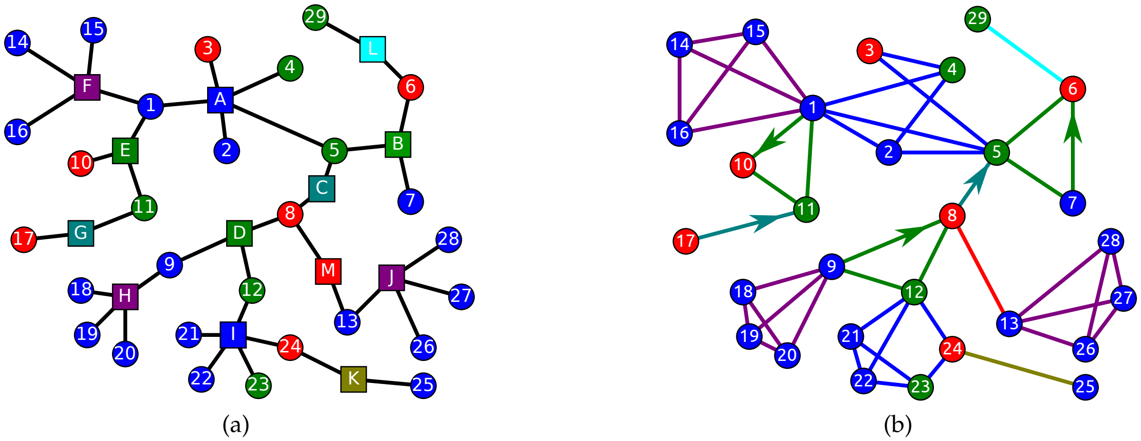


FIGURE 3.1 – Illustration of the projection process introduced in section 3.4. (a) In unclustered multitype bipartite graphs, nodes (29 circles) belong to different types (colours, $M = 3$) and are linked exclusively to groups (13 squares), which are distinguished through types as well (colours, $\Lambda = 7$). (b) In clustered multitype graphs, nodes linked to a same group in the underlying unclustered multitype bipartite graph are linked to one another through a motif whose nature and structure are specified by the corresponding group type. Labels have been added to nodes and groups for the sake of comparison between (a) and (b) and are not part of the model.

same number of type-ν groups in the unclustered multitype bipartite graph ensemble.

3.4.2 Clustered multitype graph ensemble

A clustered graph ensemble is obtained from the unclustered multitype bipartite graph ensemble by means of a projection similar to the one-mode projection of [138]. This projection is achieved by replacing the group nodes in the unclustered multitype bipartite graphs by motifs involving the nodes that were linked to a same group. The nature, either quenched (fixed) or annealed (random), and the structure of these motifs is prescribed by the corresponding group type. The resulting graphs then consist of different motifs embedded in a tree-like backbone.

Figure 3.1b illustrates a resulting clustered multitype graph where every group is replaced by a multitype quenched motif. For instance, type-*green* groups (B, D and E) are replaced by a *triangle* composed of one node of each of the $M = 3$ possible types, and whose edges are undirected except for the one between the type-*blue* node and type-*red* one that is directed. Single edges — directed (C and G) or undirected (K, L, and M) — can also correspond to motifs simply composed of two nodes. The type of each of the two nodes and the direction of the edge is prescribed by the type of the group. An example of the use of annealed motifs, where edges exist with a given probablity rather than being *a priori* fixed, is given in section 3.6.

Bond percolation can exactly be solved for the CM and its numerous variants because graphs in these ensembles have an underlying tree-like structure. Thus to take advantage of the tree-like backbone of the clustered graph ensemble, the outcome of bond percolation must be solved beforehand for each motif appearing in the graph ensemble. This solution is encoded in $Q_{i\nu}(\boldsymbol{l}|\boldsymbol{n})$ giving the probability that $\boldsymbol{l}$ nodes (i.e., l_j type-j nodes, for all j) will eventually be reached from an initial type-i node by following existing edges in a type-ν motif of size $\boldsymbol{n}$. In other words, this distribution prescribes the number of nodes (and their type) from which a given motif can be left while navigating on a graph of the clustered ensemble. It therefore "restores" the tree-like structure of the unclustered multitype bipartite graphs while retaining the effect of the clustered motifs. It is this correspondence that allows the derivation of a PGF-formalism which exactly solves the bond percolation properties of the clustered multitype graph ensemble.

In principle, a wide variety of motifs can be incorporated in our model; this variety is only limited by our ability to solve the bond percolation outcome on these motifs. Motifs can be chosen to reproduce recurring patterns of interactions found in real complex networks [128], to account for local clustering in realistic synthetic networks (see section 3.6), or for theoretical investigations (see section 3.7.5). Following the results of [7], we give in 3.9 a general method to calculate $Q_{i\nu}(\boldsymbol{l}|\boldsymbol{n})$ for most, if not all, imaginable motifs of reasonable size (the limits of the method are discussed in [7]). This method can handle quenched (fixed structure) or annealed (random structure) motifs in which edges may be directed or not. Also, nodes may belong to types which permits to model (dis-)assortative mixing and heterogeneous bond percolation [10, 7].

3.5 Bond percolation properties

We now introduce a PGF-based mathematical formalism to calculate the percolation properties of the correlated and clustered graph ensemble defined in the last section. Since PGF-based percolation formalisms have become fairly standard, the unfamiliar reader should consult recent reviews on complex network modeling (see for example [139] and references therein) for further details.

We first define $\theta_{i\nu}(\boldsymbol{x})$ as the function generating the distributions $\{Q_{i\nu}(\boldsymbol{l}|\boldsymbol{n})\}$ of the outcome of bond percolation, from an initial type-i node, on the motifs corresponding to type-ν groups. As type-ν groups may not all have the same composition (e.g., household size distribution in social networks), $\theta_{i\nu}(\boldsymbol{x})$ is calculated according to

$$\theta_{i\nu}(\boldsymbol{x}) = \sum_{\boldsymbol{n}} \frac{n_i R_\nu(\boldsymbol{n})}{\langle n_i \rangle_{R_\nu}} \left[\sum_{\boldsymbol{l}=\boldsymbol{\delta_i}}^{\boldsymbol{n}} Q_{i\nu}(\boldsymbol{l}|\boldsymbol{n}) \prod_j x_{\nu j}^{l_j - \delta_{ij}} \right] , \tag{3.2}$$

with $\boldsymbol{\delta_i} \equiv (\delta_{i1}, \ldots, \delta_{iM})$ where δ_{il} is Kronecker's delta. In equation 3.2, we average over $\frac{n_i R_\nu(\boldsymbol{n})}{\langle n_i \rangle_{R_\nu}}$ instead of over $R_\nu(\boldsymbol{n})$ to account for the fact that groups containing n_i type-i nodes are n_i times more likely to be reached from any type-i node than groups containing only one type-i node. Although equation 3.2 is not explicitly labelled in this respect, more than one motifs may be associated with a given group type. In such case, the distribution $R_\nu(\boldsymbol{n})$ gives the probability of occurence of each motif, and the left-hand sum in equation 3.2 is taken over each possible motif for which a distinct distribution $Q_{i\nu}(\boldsymbol{l}|\boldsymbol{n})$ is obtained with the method outlined in 3.9.

The function $\theta_{i\nu}(\boldsymbol{x})$ is the mathematical implementation of the correspondence between the unclustered and clustered graph ensembles discussed at the end of the last section. By generating the distribution of $\nu \to j$ edges (i.e., stemming from a type-ν group and leading to a type-j node) reached by a type-i node, one can then navigate on a unclustered multitype bipartite graph as if one were on a clustered multitype graph.

We define $g_i(\boldsymbol{x})$ as the PGF generating the distribution of the number of $\nu \to j$ edges emerging from a type-i node (i.e., emerging from the groups a type-i node is connected to)

$$g_i(\boldsymbol{x}) = \sum_{\boldsymbol{k}} P_i(\boldsymbol{k}) \prod_{\nu} \left[\theta_{i\nu}(\boldsymbol{x})\right]^{k_\nu} . \tag{3.3}$$

It is also convenient to define a PGF that generates the distribution of the number of $\nu \to j$ edges emerging from a type-i node which has itself been reached via a $\mu \to i$ edge

$$f_{\mu i}(\boldsymbol{x}) = \sum_{\boldsymbol{k}} \frac{k_\mu P_i(\boldsymbol{k})}{\langle k_\mu \rangle_{P_i}} \prod_{\nu} \left[\theta_{i\nu}(\boldsymbol{x})\right]^{k_\nu - \delta_{\mu\nu}} . \tag{3.4}$$

The averaging term used in equation 3.4 is motivated by the same argument as the one in equation 3.2. With these two PGFs, we may now compute the percolation properties of clustered multitype graph ensemble.

3.5.1 Phase transition

As a class of random graphs, clustered multitype graphs undergo a phase transition corresponding to the emergence of an extensive connected "giant" component. To locate the phase transition, let us define $\xi_{\nu j}(s)$ as the average number of $\nu \to j$ edges at a distance s from any node in any graphs of the ensemble. Due to the tree-like structure of the underlying unclustered multitype bipartite graph, each $\xi_{\nu j}(s)$ is a linear combination of all $\xi_{\nu j}(s-1)$ at distance $s-1$:

$$\xi_{\nu j}(s) = \sum_{\mu i} \left[\frac{\partial f_{\mu i}(\boldsymbol{x})}{\partial x_{\nu j}}\right]_{\boldsymbol{x}=\boldsymbol{1}} \xi_{\mu i}(s-1) \tag{3.5}$$

where

$$\frac{\partial f_{\mu i}(\boldsymbol{1})}{\partial x_{\nu j}} = \sum_{\boldsymbol{k}} \frac{(k_\nu - \delta_{\mu\nu}) k_\mu P_i(\boldsymbol{k})}{\langle k_\mu \rangle_{P_i}} \frac{\partial \theta_{i\nu}(\boldsymbol{1})}{\partial x_{\nu j}} \tag{3.6}$$

is the average number of $\nu \to j$ edges emerging from a type-i node that has been reached via a $\mu \to i$ edge. In vector notation, equation 3.5 becomes

$$\boldsymbol{\xi}(s) = \mathbf{B}\, \boldsymbol{\xi}(s-1) . \tag{3.7}$$

We see from equation 3.7 that, in general, every $\xi_{\nu j}(s)$ vanishes with increasing s if all eigenvalues of the $(M\Lambda) \times (M\Lambda)$ matrix $\mathbf{B}$ are below 1. Thus the phase transition happens when the largest eigenvalue of $\mathbf{B}$ reaches unity [3].

3.5.2 Giant Component

As there may be directed edges in the graphs (through the motifs), the giant component may have a "bow-tie" structure [147, 10]. This implies that the probability $\mathcal{P}$ of reaching the giant component may not be equal to its relative size $\mathcal{S}$. Both quantities must therefore be computed separately.

Let us define $a_{\mu i}$ as the probability that a $\mu \to i$ edge does not lead to the giant component. Because of the tree-like structure of finite components in the unclustered multitype bipartite graph, we see that $a_{\mu i}$ must satisfy the self-consistency relation

$$a_{\mu i} = f_{\mu i}(\boldsymbol{a}) \ . \tag{3.8}$$

That is, every edge reached from an edge that is not leading to the giant component must not lead to the giant component either. The probability that any type-i node does lead to the giant component is therefore given by $\mathcal{P}_i \equiv 1 - g_i(\boldsymbol{a})$, and, averaging over the node type distribution $\{w_i\}$, the probability $\mathcal{P}$ that a randomly chosen node leads to the giant component is

$$\mathcal{P} = \sum_i w_i \mathcal{P}_i = 1 - \sum_i w_i g_i(\boldsymbol{a}) \ . \tag{3.9}$$

To obtain the size of the giant component, we must calculate the probability that a given node cannot be reached from any node in the giant component. This is equivalent to computing the probability that this node does not lead to the giant component when edges are followed in the reverse direction [147, 10]. Edges in the underlying unclustered multitype bipartite graph being undirected, only $\theta_{\nu i}(\boldsymbol{x})$ needs to be modified. For instance, this can be achieved by using p_{sr} instead of p_{rs} in equation 3.17. We denote this new PGF $\bar{\theta}_{\nu i}(\boldsymbol{x})$ and we will add a bar ($\bar{\ }$) over every PGF using $\bar{\theta}_{\nu i}(\boldsymbol{x})$ instead of $\theta_{\nu i}(\boldsymbol{x})$.

Following a similar approach as for computing $\mathcal{P}$, we define $\bar{a}_{\mu i}$ as the probability that a type-i node cannot be reached from the giant component via a $\mu \to i$ edge. That is, $\bar{a}_{\mu i}$ is the probability that a neighbour of a type-i node in a type-μ group is not part of the giant component. Self-consistency then requires for $\bar{a}_{\mu i}$ to satisfy

$$\bar{a}_{\mu i} = \bar{f}_{\mu i}(\bar{\boldsymbol{a}}) \ . \tag{3.10}$$

[3] We see from equation 3.5 that $\mathbf{B}$ is a non-negative and, in general, irreducible matrix. Thus the Perron-Frobenius theorem [123] ensures that the largest eigenvalue of $\mathbf{B}$ is simple, real and positive. Moreover, the associated eigenvector is the only nonnegative eigenvector of $\mathbf{B}$.

The probability that any given type-i node is not part of the giant component is therefore $\bar{g}_i(\bar{\boldsymbol{a}})$. Considering that a fraction w_i of the nodes are of type i, the fraction of the graph occupied by type-i nodes in the giant component is

$$\mathcal{S}_i = w_i \big[1 - \bar{g}_i(\bar{\boldsymbol{a}}) \big] \ , \tag{3.11}$$

and the relative size of the giant component is

$$\mathcal{S} = \sum_i \mathcal{S}_i = 1 - \sum_i w_i \bar{g}_i(\bar{\boldsymbol{a}}) \ . \tag{3.12}$$

3.5.3 Distribution of the composition of small components

To calculate the distribution of the number of nodes of each type expected in small components, we define the PGF $A_{\mu i}(\boldsymbol{x})$ that generates the distribution of the number of edges of each type (i.e., $\nu \to j$ for all ν and j) that are *ahead* of a $\mu \to i$ edge in small components. In the large system size limit, the small components have a tree-like structure and no finite-size effects are to be expected [i.e., the joint degree distribution $P_i(\boldsymbol{k})$ is constant]. We therefore expect $A_{\mu i}(\boldsymbol{x})$ to be invariant under translation on a small component; the distribution of the number of each edge type ahead, $A_{\mu i}(\boldsymbol{x})$, is independent of the position in a small component. This implies that $A_{\mu i}(\boldsymbol{x})$ must satisfy

$$A_{\mu i}(\boldsymbol{x}) = x_{\mu i} f_{\mu i}\big(\boldsymbol{A}(\boldsymbol{x})\big) \tag{3.13}$$

where the extra $x_{\mu i}$ accounts for the $\mu \to i$ edge that has just been followed. This extra factor guarantees that a finite extent of the distribution generated by $A_{\mu i}(\boldsymbol{x})$ can be obtained in a finite number of iterations of equation (3.13) starting with the initial conditions $A_{\mu i}(\boldsymbol{x}) = 1$. Replacing $x_{\nu i} = z_i$ for all ν in $A_{\mu i}(\boldsymbol{x})$ generates the distribution of the number of nodes of each type ahead a type-i nodes reached from a type-μ group. Thus the composition of a small component reached from a type-i node is generated by $z_i g_i\big(\boldsymbol{A}(\boldsymbol{z})\big)$; again the extra z_i accounts for the initial type-i node. Because any node is of type i with probability w_i, the composition of a small component that is reached from a randomly chosen node is therefore generated by

$$K(\boldsymbol{z}) = \sum_i \frac{w_i z_i g_i\big(\boldsymbol{A}(\boldsymbol{z})\big)}{1 - \mathcal{P}} \ , \tag{3.14}$$

where $1 - \mathcal{P}$ ensures the normalization of $K(\boldsymbol{z})$. Note that $A_{\mu i}(\boldsymbol{1})$ is equal to the probability that a $\mu \to i$ edge leads to a finite (small) component, and is therefore equal to $a_{\mu i}$.

Solving equations (3.13)–(3.14) can however become tedious when dealing with large number of types of nodes and groups, or large groups. It is therefore worth noting that the first moments of the distribution generated by $K(\boldsymbol{z})$ can be calculated in a more direct manner. For instance, let us compute the average number $\langle s_i \rangle$ of type-i nodes in small components. With

$$\langle s_i \rangle = \left. \frac{\partial K(\boldsymbol{z})}{\partial z_i} \right|_{\boldsymbol{z}=\boldsymbol{1}}$$

inserted in equation (3.14), replacing $x_{\mu i}$ with z_i, we get

$$\langle s_i \rangle = \frac{w_i(1-\mathcal{P}_i)}{1-\mathcal{P}} + \sum_{j\gamma r} \frac{w_j \langle k_\gamma \rangle_{P_j} a_{\gamma j}}{1-\mathcal{P}} \frac{\partial \theta_{j\gamma}(\boldsymbol{a})}{\partial x_{\gamma r}} \frac{\partial A_{\gamma r}(\mathbf{1})}{\partial z_i} \ , \tag{3.15}$$

where we have used equation (3.4), equation (3.8) and the fact that $g_i(\boldsymbol{a}) = 1 - \mathcal{P}_i$. In this last result, $\frac{\partial \theta_{j\gamma}(\boldsymbol{a})}{\partial x_{\gamma r}}$ is the average number of type-r nodes that are accessible from a type-j node in a type-γ group in small components. Also, $\frac{\partial A_{\gamma r}(\mathbf{1})}{\partial z_i}$ is the average number of type-i nodes ahead of a $\gamma \to r$ edge in small components. From equation (3.8), we see that this last quantity is the solution of

$$\frac{\partial A_{\gamma r}(\mathbf{1})}{\partial z_i} = a_{\gamma r}\delta_{ir} + \sum_{\lambda s} \frac{\partial f_{\gamma r}(\boldsymbol{a})}{\partial \theta_{r\lambda}} \frac{\partial \theta_{r\lambda}(\boldsymbol{a})}{\partial x_{\lambda s}} \frac{\partial A_{\lambda s}(\mathbf{1})}{\partial z_i} \tag{3.16}$$

where $\frac{\partial f_{\gamma r}(\boldsymbol{a})}{\partial \theta_{r\lambda}}$ is the average number of type-λ groups to which a type-r node reached via a type-γ group is connected in small components. Thus by solving equations (3.8)–(3.9) and then equations (3.15)–(3.16), it is possible to obtain quite easily the average number of nodes of each type in the small components. Equations for higher moments can be obtained in a similar manner and are straightforward to derive.

3.6 Illustration and validation

To illustrate the versatility and the usefulness of our approach, we generated *urban networks* [126] and used our formalism to predict the outcome of an outbreak of a hypothetical infectious disease. In these graphs, three ($M = 3$) types of nodes – namely adults (type 1), heath-care workers (HCW, type 2) and children (type 3) – interact within groups representing households, workplaces, schools and hospitals. In addition, friendship bonds between children are modeled using a nontrivial motif (shown in figure 3.4 in 3.10), and directed edges from adults and children to HCW are added to account for the susceptibility of HCW to get infected by people seeking care in hospitals [14]. The disease spreads from infectious nodes to their neighbours with probability T called the *transmissibility* [136]. Further details of these graphs and of the associated numerical simulations are relegated to 3.10. It should be appreciated that these graphs contain a wide range of properties found in real complex networks such as clustering of several orders (e.g., arbitrary motifs, heterogeneous Erdős-Rényi cliques), (dis)assortative mixing, degree-degree correlation and directed edges.

Figure 3.2 shows the typical bifurcation diagram of the giant-component-related quantities $\mathcal{P}$ and $\{\mathcal{S}_i\}$. Apart from the excellent agreement between the results of the numerical simulations and the predictions of our formalism, this figure illustrates how the multitype approach can highlight the behavioral differences between different populations — identified by their own node type — within a same graph ensemble. In this specific case, the HCW population has purposely been put in the situation where each HCW has more incoming edges than outgoing edges with adults and children. Also, the average degree inside the

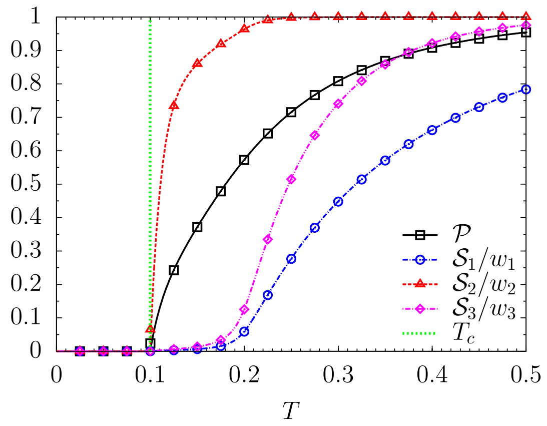
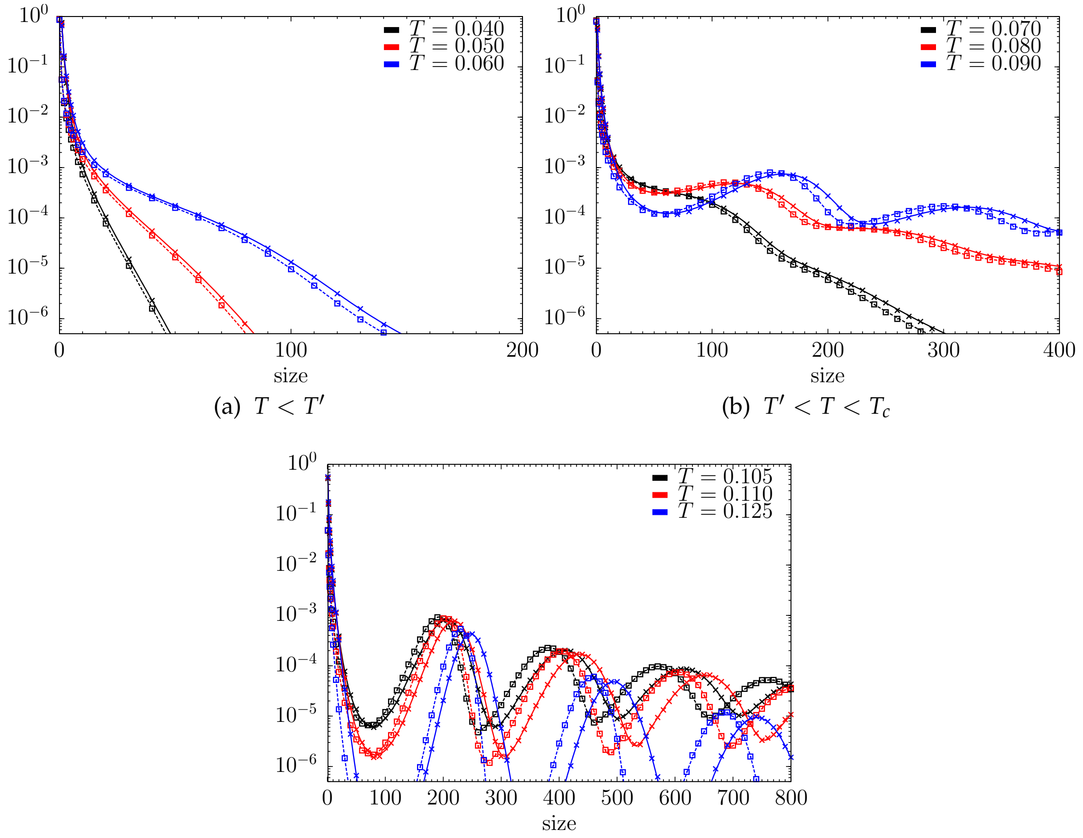


FIGURE 3.2 – Bifurcation diagram of the probability to reach the giant component $\mathcal{P}$ and the fraction of nodes of type i therein $\mathcal{S}_i/w_i$ as a function of the occupation probability of edges (or transmissibility) T. Types 1, 2 and 3 correspond respectively to adults, HCW and children. Lines represent the theoretical predictions of our formalism [equations (3.8)–(3.11)] while symbols have been obtained through numerical simulations (over 10^5 simulations on graphs of at least 1.2×10^5 nodes for each symbol, see 3.10 for further details). The percolation threshold $T_c \simeq 0.1$ has been obtained by finding the value of T for which the largest eigenvalue of $\mathbf{B}$ equals 1 [see equation (3.7)].

Erdős-Rényi cliques corresponding to hospitals (300 nodes connected to one another with probability 0.05) is greater than 1 for T greater than $T' \equiv [0.05 \times 299]^{-1} \simeq 0.067$. This implies that these cliques are increasingly likely to have percolated (i.e., to have a spanning cluster) for $T > T'$. Qualitatively, once an outbreak reaches the HCW population, it is likely to stay mostly confined in it and to infect a large proportion of it. Only when T becomes sufficiently large does the outbreak invade other part of the population (schools, workplaces and friendship circles). These insights are corroborated by figure 3.2. It also shows that although the HCW population accounts for only 5% of the total population, it drives the percolation process by pulling down its threshold to $T_c \simeq 0.1$; the other node types only significantly join (i.e., $\mathcal{S}_i/w_i > 0.01$) the giant component at $T \simeq 0.14$ and $T \simeq 0.16$, respectively.

Figure 3.3 shows the distribution of the total number of nodes in small components for various values of T. To support our claim that outbreaks are mostly confined within the HCW populations, figure 3.3 also displays the distribution of the number of nodes of type 2 in the small components. The small shift between the two curves is due to adults and children being infected mostly in households. Again, we conclude in an excellent agreement between both the numerical simulations and theoretical predictions of the formalism obtained by solving equations (3.13)–(3.14).

Interestingly, figures 3.3a–3.3c give evidence of what one may call the "local percolation" of the hospital cliques as T increases. For $T < T'$, the size distribution falls rapidly and



(a) $T < T'$

(b) $T' < T < T_c$

(c) $T > T_c$

FIGURE 3.3 – Distribution of the number of nodes in small components for various values of the transmissibility T (one colour per value). Continuous and dashed curves represent the total number of nodes and the number of type-2 nodes (HCW) in the small components, respectively. Lines were obtained by solving equations (3.13)–(3.14) and symbols were obtained through numerical simulations (over 10^8 simulations on graphs of at least 4.8×10^5 nodes for each symbol, see 3.10 for further details).

monotonously as expected for generic CM graphs [147, 142]. For $T' < T < T_c$, however, the shape of the distribution changes as local maxima appear. These are due to the growing spanning cluster in the hospital cliques. For $T > T_c$, most of the HCW population is part of the giant component, and the spanning cluster is more and more likely to cover the entire clique as T increases. The HCW nodes that are not part of the giant component are therefore likely to be part of very large small components composed of one or more "locally percolated" cliques. This is confirmed by the multiple maxima seen on figures 3.3b–3.3c.

3.7 Special cases, generalization and applications

We now demonstrate our claims that our formalism encompasses many percolation models on random graphs published in the litterature. We also succinctly outline a possible generalization and some straightforward applications of our model.

3.7.1 Multitype random graphs

Our formalism naturally falls back on the model introduced in [10] describing the heterogeneous bond percolation on multitype random graphs. In this class of graphs, there are M types of nodes, and a $i \rightarrow j$ edge is occupied with probability T_{ij}. Type-i nodes occupy a fraction w_i of the graph, and a type-i node is connected to $\tilde{k}_j$ type-j nodes (for each $j \in [1, M]$) with probability $\tilde{P}_i(\tilde{k}_1, \tilde{k}_2, \ldots, \tilde{k}_M)$.

Our formalism reproduces this model by using one group type for each possible (unordered) type of $i \rightarrow j$ edge. To each of the $\Lambda = M(M+1)/2$ group types are associated the functions

$$\begin{aligned}\theta_{i\nu}(\boldsymbol{x}) &= [1 + (x_{\nu j} - 1)T_{ij}] \\ \theta_{j\nu}(\boldsymbol{x}) &= [1 + (x_{\nu i} - 1)T_{ji}]\end{aligned}$$

depending whether the edge is considered in the $i \rightarrow j$ or in the $j \rightarrow i$ direction. Along with these functions, $P_i(\boldsymbol{k})$ can therefore reproduce the degree distribution $\tilde{P}_i(\tilde{k}_1, \tilde{k}_2, \ldots, \tilde{k}_M)$.

As shown in [10], multitype random graphs naturally encompasses multipartite graphs, as well as the undirected random graphs introduced in [147, 136, 138]. By assigning nodes with a given degree to a same node type, our formalism can also reproduce degree-degree correlation as in [135].

3.7.2 Clustered random graphs

Being a multitype generalization of the *highly clustered random graphs* introduced in [138], our model simplifies to the latter in a straightforward manner with $M = \Lambda = 1$ and all groups being Erdős-Rényi cliques. For $\Lambda = 1$, the groups to which any given node belongs to is averaged in equation (3.2) so that no correlation whasoever can be taken into account.

When considering only $M = 1$ type of nodes but an arbitrary number of uniquely configured groups [$R_\nu(\boldsymbol{n}) = 1$ for all ν], we retrieve *random graphs containing arbitrary distributions of subgraphs* as introduced in [98]. The unweighted average equation (3.19) plays an analogous function as their *role* distribution, with correlation being taken into account by using node types. It is then straightforward to conclude that our formalism also encompasses the *edge-triangle* model introduced in [127, 143] and the *strong ties* model proposed by [180].

The γ*-theory* model [80] can be recovered by considering only $M = 1$ type of nodes, and by allowing nodes to belong to only one group of size larger than two (Erdős-Rényi cliques) but to belong to an arbitrary number of group of size two (external edges). Also, the *random hypergraphs* introduced in [77] can be reproduced by our formalism by considering $M = 3$ types of nodes and $\Lambda = 1$ type of groups which are triangles composed of one node of each type.

Finally, a class of formalism [177, 202, 176] uses the *multiplicity* of edges — the number of triangles to which an edge participates — to derive an effective branching process and solve the percolation on clustered graphs using PGFs. Although this approach tackles percolation from a different perspective, its predictions (*i.e.*, the percolation threshold and the size of the giant component) can be reproduced with our model by using fully connected motifs of size $m + 2$ to account for links of multiplicity m and by appropriately using node and group types to account for the correlations that this class of models incorporates.

3.7.3 Directed random graphs

Our formalism as presented in this paper can only model directed edges between *different* node types. To describe directed edges among a same node type such as in [147, 125], we would need to subdivide the group type corresponding to directed edges into an incoming part and an outgoing part (e.g., $\nu \to \nu_{\text{in}}, \nu_{\text{out}}$), and match the complementary parts to form groups in the unclustered multitype bipartite graph. In other words, each group is linked to an incoming and an outgoing stub. The mathematical formalism introduced in section 3.5 remains valid except that we would need to explicitly consider the fact that nodes are reached by their incoming edges and are left by their outgoing edges when writing down the equations (see [147, 125] for detailed examples). This adjustment is nevertheless straightforward and does not affect the generality of our approach.

3.7.4 Interdependent or coupled networks

The use of node and group types naturally permits our formalism to be used in the study of interdependent or coupled networks. In interdependent – or interacting – networks, node types could for instance be used to distinguish the elements of two or more interacting networks [108, 34, 95]. Different group types would then allow to specify precisely the (non-trivial) interactions within and across the networks. In the case of coupled, or overlayed, networks [117, 72] elements in a single population interact in different ways which is modelled using different edge types. This again can be easily achieved with our formalism by defining multiple group types, one for each level of interaction. Again, the generality of our approach gives us access to a wide variety of complex patterns of interactions in a very detailed fashion.

3.7.5 Weak and strong clustering regimes

The existence of two regimes of clustering, *weak* and *strong*, has been put forward in [177, 175, 176] with the conclusion that these two regimes have opposite effect on the bond percolation threshold. In the weak regime, edges have a multiplicity of either 0 or 1 (single edges or disjoint triangles), and the percolation threshold is higher than for equivalent unclustered graphs. In the strong regime, edges may contribute to more than one triangle, and it is argued that the percolation threshold is then lower than for equivalent unclustered graphs.

Contrariwise, the analysis done in [127, 80] strongly suggests that clustering always increases the percolation threshold and that the observed lower percolation threshold in the strong regime is due to assortative mixing instead. Hence, according to theses results, there should be no weak and strong clustering regimes. The use of node and group (or edge) types in our model can generate clustered and unclustered graphs with the same correlations (or mixing patterns). It is therefore possible to investigate — both numerically and analytically — the effect of clustering alone on the percolation threshold, shedding some light on this contradiction while extending the analysis and the conclusions of [127, 80]. This will be addressed in a future publication.

3.8 Conclusion

We have presented a generalization of the Configuration Model allowing for the inclusion of several nontrivial mixing patterns and clustering. On the one hand, the use of node and group types permits to explicitly prescribe how nodes are connected to one another, hence reproducing (dis-)assortative mixing, and indirectly degree-degree correlation. On the other hand, the use of a one-mode projection can generate a wide range of nontrivial clustered structures through quenched or annealed motifs. Besides the modeling of mixing patterns, the multitype approach permits to identify nodes. This allows to highlight unusual behaviors or susceptibility of sub-population of nodes, as well as to simulate targetted intervention such as attacks, failures, vaccination or quarantine. We have also demonstrated that our formalism encompasses several models published to this day, and we have outlined potential applications.

Bridging the gap between empirical network datasets and theoretical models is surely one the principal tenets of network theory. Since extracting the effective clustered backbone (*i.e.*, motifs) of real networks is still an open problem, our approach can only offer a partial answer. However, it provides a comprehensive synthesis of the many variants of the CM published to date, and it extends considerably the structural complexity of graphs that can be handled theoretically. In these regards, the versatility and generality of the present framework could prove useful even beyond the strict confines of bond percolation on complex graphs.

3.9 Appendix: General method to solve bond percolation on arbitrary multitype motifs

We present a systematic way to compute the outcome of bond percolation, $Q_{i\nu}(\boldsymbol{l}|\boldsymbol{n})$, on any arbitrary multitype motifs where edges are simple and can be directed or not.

Let us first consider a multitype generalization of Erdős-Rényi random graphs. These are composed of $\boldsymbol{n}$ nodes, and a directed edge exists from a type-i node to a type-j node with probability p_{ij}. Edges exist independently of one another. Note that the symmetric case $p_{ij} = p_{ji}$ is statistically equivalent to undirected edges. It has been shown [7] that $Q_{i\nu}(\boldsymbol{l}|\boldsymbol{n})$ can be obtained by iterating

$$Q_{i\nu}(\boldsymbol{l}|\boldsymbol{n}) = Q_{i\nu}(\boldsymbol{l}|\boldsymbol{l}) \prod_{rs} \binom{n_r - \delta_{ir}}{l_r - \delta_{ir}} (1 - p_{rs})^{l_r(n_s - l_s)} \tag{3.17}$$

and

$$Q_{i\nu}(\boldsymbol{l}|\boldsymbol{l}) = 1 - \sum_{\boldsymbol{m}<\boldsymbol{l}} Q_{i\nu}(\boldsymbol{m}|\boldsymbol{l}) \tag{3.18}$$

from the initial condition $Q_{i\nu}(\boldsymbol{\delta_i}|\boldsymbol{\delta_i}) = 1$ with $\boldsymbol{\delta_i} \equiv (\delta_{i1}, \ldots, \delta_{iM})$. In essence, knowing the probability of finding a component of size $\boldsymbol{l}$ from a node of type i in a graph of size $\boldsymbol{l}$, equation (3.17) computes the probability of finding a sub-component of size $\boldsymbol{l}$ but in a graph of size $\boldsymbol{n}$ ($> \boldsymbol{l}$). This allows to compute every coefficients of the distribution $Q_{i\nu}(\boldsymbol{l}|\boldsymbol{n})$ except for the last one, the one corresponding to the case where the whole graph is reachable, which is obtained using equation (3.18).

Let $\mathcal{G}$ be a multitype motif composed of $\boldsymbol{n}$ nodes with an arbitrary configuration of edges. The associated distribution $Q_{i\nu}(\boldsymbol{l}|\boldsymbol{n})$ can then be computed by following these simple steps:

1. Consider an equivalent multitype Erdős-Rényi graph $\mathcal{G}'$ of size $n' = \sum_j n_j$ in which each node belongs to its own unique type (i.e., $n'_{j'} = 1$ for all $j' \in \{1, \ldots, n'\}$). Note $p'_{i'j'}$ the probability for a directed edge to exist from the type-i' node to the type-j' one.
2. Compute $Q'_{i'\nu}(\boldsymbol{l'}|\boldsymbol{n'})$ for $\mathcal{G}'$ with equations (3.17)–(3.18). Without any loss of generality suppose that the initial node from which the graph is probed is of type 1 [i.e., equations (3.17)–(3.18) need to be solved only once].
3. From $Q'_{i'\nu}(\boldsymbol{l'}|\boldsymbol{n'})$, derive the intermediate distribution $Q^{(j)}_{i\nu}(\boldsymbol{l}|\boldsymbol{n})$ of the number of nodes of each type that are accessible from the j-th type-i node. This is achieved by replacing the *artificial* node types in $\mathcal{G}'$ by the *actual* node types in $\mathcal{G}$, and by setting the values of $p'_{i'j'}$ according to the configuration of the edges in $\mathcal{G}$, which can include type-dependent probabilities of existence/occupation of edges.
4. Obtain $Q_{i\nu}(\boldsymbol{l}|\boldsymbol{n})$ by computing the unweighted average of the n_i distributions $Q^{(j)}_{i\nu}(\boldsymbol{l}|\boldsymbol{n})$

$$Q_{i\nu}(\boldsymbol{l}|\boldsymbol{n}) = \frac{1}{n_i} \sum_j Q^{(j)}_{i\nu}(\boldsymbol{l}|\boldsymbol{n}) \,. \tag{3.19}$$

Group type	Composition $\boldsymbol{n} = (n_1, n_2, n_3)$	Probability $R_\nu(\boldsymbol{n})$
Households	$(2, 0, 0)$	0.0810
	$(1, 1, 0)$	0.0180
	$(0, 2, 0)$	0.0010
	$(2, 0, 1)$	0.1215
	$(1, 1, 1)$	0.0270
	$(0, 2, 1)$	0.0015
	$(2, 0, 2)$	0.3240
	$(1, 1, 2)$	0.0720
	$(0, 2, 2)$	0.0040
	$(2, 0, 3)$	0.2835
	$(1, 1, 3)$	0.0630
	$(0, 2, 3)$	0.0035
Schools	$(5, 0, 50)$	0.2500
	$(10, 0, 100)$	0.5000
	$(15, 0, 150)$	0.2500
Workplaces	$(10, 0, 0)$	0.1000
	$(20, 0, 0)$	0.2500
	$(30, 0, 0)$	0.3000
	$(40, 0, 0)$	0.2500
	$(50, 0, 0)$	0.1000
Hospitals	(0,300,0)	1.0000
Friendships	(0,0,5)	1.0000
Directed edges (1 → 2)	(1,1,0)	1.0000
Directed edges (3 → 2)	(0,1,1)	1.0000

TABLE 3.2 – Distribution $R_\nu(\boldsymbol{n})$ used for the simulations in section 3.6 with $M = 3$ and $\Lambda = 7$.

A noteworthy point is that the distribution $Q'_{i\nu}(\boldsymbol{l'}|\boldsymbol{n'})$ computed for a generic graph of size n' can generate every multitype motif of size smaller than n' by appropriate choices of $p'_{i'j'}$. An explicit example of such a calculation is given in [7].

3.10 Appendix: Numerical simulations

Details of the graphs used in section 3.6 and of the numerical simulations performed to validate our formalism are presented.

3.10.1 Urban networks

The graphs generated in section 3.6 were inspired by the *urban networks* used in [126, 14] in which individuals are connected to one another because of their common membership to a social group (e.g., households, schools, workplaces, hospitals, friendship circles). In this case, the population is divided into three categories – identified by node types – namely adults

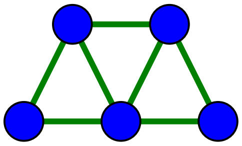


FIGURE 3.4 – Motif used to model friendship bonds between children in the *urban network* used in section 3.6.

(type 1), health-care workers (HCW, type 2) and children (type 3) with $\{w_i\} = \{0.45, 0.05, 0.50\}$. Every node belongs to one household, every HCW belongs to one hospital, every child belongs to one school, 1/9 of adults belong to one school (teachers, janitors, etc.) and the remaining 8/9 belong to one workplace. Also every child belongs to one group of friends (see figure 3.4), and adults and children are connected at most to two randomly chosen HCW via a directed edge.

Table 3.2 explicits the group composition distribution $R_\nu(\boldsymbol{n})$ used to generate the urban networks. Except for friendship circles, the connections between individuals within groups are modeled with multitype Erdős-Rényi graphs with different probabilities of edge existence. In households, every possible edge exists except for the directed edges from HCW to adults, HCW and children that exist with probability 0.2, 0.2 and 0.1, respectively. In schools and workplaces, edges exist with probability 0.01 and they exist with probability 0.05 in hospitals. The use of relatively large cliques with such low probabilities of existence of edges allows to model redundancy in the neighbourhood of nodes while keeping a relatively low clustering. Finally, directed edges from adults and children to HCW exist with probability 0.5.

These graphs can be generated in a fairly straightforward manner. For a given group type ν, we first generate a sequence of groups whose composition is prescribed by $R_\nu(\boldsymbol{n})$. We then generate, according to $P_i(\boldsymbol{k})$, a list of nodes in which a node belonging to k_ν type-ν groups appears k_ν times. We finally randomly assign these nodes to the groups, and create edges between nodes that are members of a same group according to the probabilities given in the last paragraph.

3.10.2 Percolation simulations

Graphs that were used to obtain the results shown in section 3.6 were composed of at least $1.2{\times}10^5$ nodes. For T around T_c, larger graphs (up to $9.6{\times}10^6$ nodes) have been generated to faciliate the distinction between small components, which are intensive, from the giant component, which is extensive. At least 10^3 (10^6) graphs were generated for each value of T used in figure 3.2 (figure 3.3).

For each generated graph, 100 percolation simulations were performed. These consist in randomly choosing a starting node and then following every possible edges leaving this node – and the subsequently encountered nodes – with probability T until no new node can be reached. The component size is then simply the number of nodes that have been reached. While it would have been straightforward to use a type-specific probability T (see section 3.7.1), we have used a single value to lighten the presentation of the results.

Chapitre 4

Théorie III : D'une approche générale et exacte à la percolation sur graphes

Article original :

A general and exact approach to percolation on random graphs

Antoine Allard[1], Laurent Hébert-Dufresne, Jean-Gabriel Young et Louis J. Dubé

Département de Physique, de Génie Physique, et d'Optique,
Université Laval, Québec (Qc), Canada G1V 0A6

Référence : Sera soumis à Phys. Rev. E

[1] Le développement théorique, les simulations numériques de même que la rédaction du manuscrit sont essentiellement les fruits du travail du premier auteur. Les autres auteurs ont participé à l'élaboration générale du projet de même qu'à la révision du manuscrit.

4.1 Avant-propos

Cet article présente l'aboutissement du travail théorique effectué au cours de cette thèse. Comme nous le laissâmes entendre dans l'avant-propos du chapitre précédent, cette nouvelle version du formalisme n'était pas prévue. Nous considérions alors la version présentée au chapitre 3 comme un point final à nos efforts de généralisation pour cette classe de modèles ; les quelques pistes de développement possibles ne nous semblaient pas avoir l'envergure suffisante pour être poursuivies. C'est toutefois en utilisant le formalisme dans le contexte des projets présentés aux annexes A et B que ces développements devinrent nécessaires et conséquemment que l'intérêt de les formaliser dans un cadre théorique unifié se manifesta.

Il s'agit de la version la plus générale de l'ensemble CM et du formalisme associé ; de fait, les outils théoriques présentés dans les autres chapitres de cette thèse n'en sont que des cas spéciaux. À nouveau, nous exploitons « l'approche multitype » : une idée banale et simple à première vue, mais pourtant extrêmement puissante et conviviale. En plus d'identifier les nœuds à l'aide de types, nous attribuons des types aux demi-liens via lesquels les nœuds sont liés les uns aux autres. Ce passage des types de groupes aux types de demi-liens augmente significativement la complexité des graphes, tout en préservant les propriétés structurelles qui pouvaient déjà être modélisées dans la version précédente du formalisme. Ceci se voit notamment par la simplicité avec laquelle les liens unidirectionnels sont dorénavant intégrés, et l'annexe A présente un modèle où les types des demi-liens indiquent une fonction particulière et non réciproque des liens (c.-à-d. la fonction du lien change selon le sens dans lequel il est parcouru). Enfin, nous montrons comment notre approche peut-être adaptée aux *graphes interdépendants* caractérisés par l'émergence d'une composante extensive via une transition de phase discontinue. Du coup, nous présentons le modèle le plus général de graphes interdépendants. Des résultats préliminaires nous permettent de conjecturer l'impact qu'a l'agrégation sur la position de la transition de phase, de même que sur l'amplitude de la discontinuité.

4.2 Abstract

We present a comprehensive and versatile theoretical framework to study site and bond percolation on clustered and correlated random graphs. Our contribution can be summarized in three points. (i) We introduce a set of iterative equations that solve exactly the distribution of the sizes of components in finite size quenched or random graphs. (ii) Our framework leads naturally to a definition of a very general random graph ensemble that encompasses most of the model published to this day, but also permits to model structural properties that were yet to be included in a theoretical framework. Site and/or bond percolation on this ensemble is solved exactly in the thermodynamical limit using probability generating functions (i.e., the percolation threshold, the sizes of the extensive component of the small components are obtained). (iii) Our approach can be adapted to model interdependent graphs—whose most striking feature is the emergence of the extensive component via a discontinuous transition—in an equally general fashion. Among other things, this provides the first theoretical framework that includes clustering in interdependent graphs, and preliminary results suggest that clustering increases the discontinuity amplitude at the transition.

4.3 Introduction

Percolation on graphs offers a simple theoretical framework to model and investigate the behavior of many complex systems; noteworthy examples being the growth and the robustness of these systems [44, 58], their observability [9, 200], as well as the propagation of emerging infectious diseases [91, 124]. On the analytical front, recent progress has been mainly achieved within the Configuration Model (CM) paradigm [144], which in the limit of large graphs allows an exact and simple analytical treatment with the use of probability generating functions (pgf) [37, 147]. The versatility of the pgf method has triggered the development of many variants of the CM reproducing, to some extent, correlations and clustering found in real complex systems [6, 10, 23, 77, 80, 84, 92, 98, 102, 108, 125, 127, 135, 137, 138, 143, 151, 177, 180, 192, 193, 202], thus deepening our understanding of the interplay between the organization and the function of complex systems.

We present a very general class of random graphs that reproduces a great variety of nontrivial correlations and clustering patterns found in real complex networks. These correlations are incorporated into the graphs through the use of types of nodes and types of stubs (i.e., half-edge stemming from nodes). Hence, by explicitly controlling *who is connected to whom* and *through what kind of connection*, our approach reproduces any correlations as long as they can be mapped unto this multitype framework. For instance, node types can correspond to the degree of nodes (the number of neighbors) [135, 193], to their intrinsic properties such as age or ethnicity [10, 137], or to their position in the k-core structure of the graph [92].

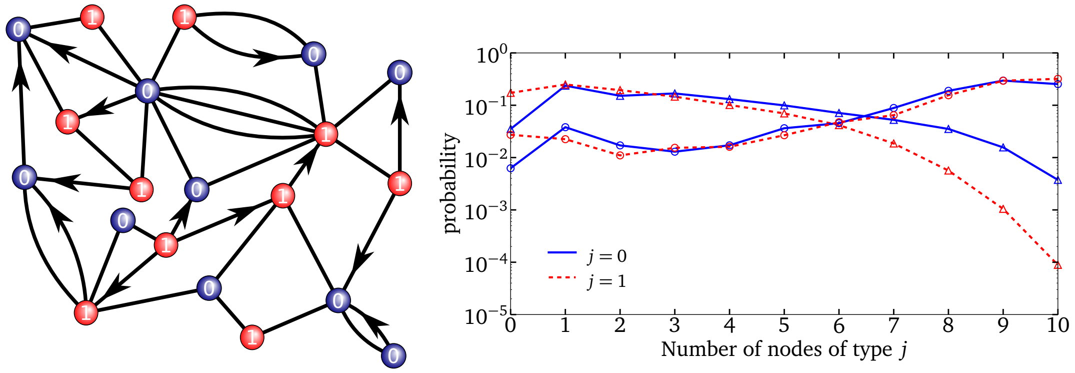
Furthermore the use of types of stubs allows to explicitly account for different categories of connections. On the one hand, these differences may be of a *conceptual* nature [105]. For instance in *multilayer* or *multiplex* networks the type of an edge refers to the layer of interaction to which it belongs (e.g., family ties and acquaintances in social networks). Also, in *interacting* or *interdependent* networks edges are distinguished according to whether they connect nodes of the same network (e.g., *connectivity links*) or connect nodes of two different networks (e.g., *dependency links*).

On the other hand, the different types of stubs can account for different *topological* functions. For instance, as some edges may be undirected or directed, different types of stubs can be used to identify *in-degrees*, *out-degrees* or *undirected degrees* [125, 147]. More importantly, stubs can be matched in groups of more than two nodes to form *motifs*, or *hyperedges* [6, 98], permitting the inclusion of clustering in a very general fashion. These motifs can take a wide variety of forms: simple triangles, cliques of several hundreds of nodes, or arbitrary graphs with directed and multiple edges [see Fig. 4.1(a)]. Additionally, these motifs can have a quenched (i.e., fixed) or random structure (e.g., Erdős-Rényi graphs).

We develop a mathematical framework that solves the site and bond percolation (hereafter referred to as *hybrid* percolation) on this general class of random graphs. We build upon the well-known pgf-based formalism and obtain in the limit of large graph size the analytical expression for the size of the extensive "giant" component, the percolation threshold, as well as the distribution of the sizes of the "small" components. The pgf approach *de facto* assumes locally tree-like graphs which forbids closed loops and therefore any clustering whatsoever. To circumvent this limitation, we present a set of iterative equations that exactly solves the size distribution of components in finite-size arbitrary, or quenched, graphs. These equations map the possible outcomes of hybrid percolation on any motifs (i.e., the size distribution of the components) unto a distribution of branching trees, and thereby reconcile the presence of motifs with the tree-like requirement of the pgf approach.

The general nature of our model acts as a *theoretical laboratory* where the effect of a wide selection of structural features on the outcomes of hybrid percolation can be investigated on a common ground. We provide several illustrations of this claim throughout this paper. Moreover, our model encompasses most, if not all, variants of the CM published to date. We provide several examples supporting this claim as well. Finally, our mathematical formalism can be adapted to describe hybrid percolation on interdependent graphs, and preliminary results suggest that clustering increases the jump of the discontinuity at the transition.

This paper is organized as follows. In Sec. 4.4, we introduce the set of iterative equations that exactly solves the size distribution of components in finite-size arbitrary graphs. In Sec. 4.5, we formally define the general graph ensemble discussed above and obtain its exact structural properties under hybrid percolation (i.e., obtain the sizes of the components and the percolation threshold). We show in this same section how our formalism can be adapted to



(a)

(b)

FIGURE 4.1 – (a) Example of an arbitrary graph that can be handled by our framework. Blue and red represent node types 0 and 1. There are 10 nodes of each type. (b) Distribution of the number of nodes of type j in components reached from an initial node chosen at random for the graph shown in (a). Symbols are the results of 2.5×10^8 numerical simulations, and lines are the predictions of Eqs. (4.1a)–(4.1c). Triangles ($\triangle$) correspond to pure site percolation with $\{\tilde{r}'_s\} \equiv \{\tilde{r}'_0, \tilde{r}'_1\} = \{0.70, 0.40\}$ and $\{\tilde{p}'_{ij}\} \equiv \{\tilde{p}'_{00}, \tilde{p}'_{01}, \tilde{p}'_{10}, \tilde{p}'_{11}\} = \{1.00, 1.00, 1.00, 1.00\}$. These probabilities are given in terms of the original node types to lighten the presentation (denoted by a prime). Consequently to use the mapping described in Sec. 4.4.2, a probability for each individual node and each individual edge must be defined. For instance, if node 2 is of type 0, we set $\tilde{r}_2 = 0.40$. Circles ($\circ$) correspond to hybrid percolation (site and bond) with $\{\tilde{r}'_s\} \equiv \{\tilde{r}'_0, \tilde{r}'_1\} = \{0.90, 0.95\}$ and $\{\tilde{p}'_{ij}\} \equiv \{\tilde{p}'_{00}, \tilde{p}'_{01}, \tilde{p}'_{10}, \tilde{p}'_{11}\} = \{0.95, 0.85, 0.90, 1.00\}$. The distributions are discrete; lines have been added to guide the eye.

interdependent graphs. We then illustrate the workings of our formalism with several examples and special cases in Sec. 4.6. Conclusions and final remarks are collected in Sec. 4.7.

4.4 Percolation on finite-size arbitrary graphs

To reconcile the tree-like assumption of the pgf approach with the presence of motifs in graphs, the outcomes of percolation on these motifs—the distribution of the number of nodes that can be reached from a given node—must be computed beforehand. These distributions can be computed by hand by enumerating each possible configuration where nodes and edges exist with given probabilities [98, 127], but doing so becomes quickly tedious for motifs of more than a handful of nodes. Instead, we generalize the equations presented in Refs. [7, 138] to obtain a set of iterative equations that solve the outcomes of hybrid percolation on small arbitrary graphs.

4.4.1 Multitype Erdős-Rényi graphs

Let us first consider multitype random graphs as a generalization of the $\mathcal{G}_{n,p}$ model (Erdős-Rényi random graphs) in which n nodes are linked by edges that exist individually and independently with a probability p [78]. We generalize this model by labeling nodes using types; the set of types is noted $\mathcal{N}$, and there are a total of $|\mathcal{N}|$ types of nodes. Furthermore, a directed edge from a type i node to a type j node (noted $i \to j$) exists with a probability p_{ij} independently of other potential edges[2]. For the sake of conciseness, we will refer to a graph composed of n_i type i nodes (with $i = 1, \dots, |\mathcal{N}|$) with the vector $\boldsymbol{n} \equiv (n_1, \dots, n_{|\mathcal{N}|})^{\mathsf{T}}$. We will use a similar notation for other quantities throughout this paper, unless specified otherwise.

Since edges may be directed, we define a component to be the nodes that are *reachable* from a given node, also included in the component. This initial node is identified solely by its type since nodes of a given type are indistinguishable. We define $W_i(\boldsymbol{l}|\boldsymbol{n})$ as the probability that $\boldsymbol{l} \equiv (l_1, \dots, l_{|\mathcal{N}|})^{\mathsf{T}}$ nodes can be reached from an initial node of type i in a graph containing $\boldsymbol{n}$ nodes. As shown in Ref. [7], the distribution $\{W_i(\boldsymbol{l}|\boldsymbol{n})\}$ can be obtained by iterating a set of equations. We briefly sketch how to derive these equations; the interested reader is referred to the original publication for more details.

The calculation of $W_i(\boldsymbol{l}|\boldsymbol{n})$ begins with the initial condition $W_i(\boldsymbol{\delta}_i|\boldsymbol{\delta}_i) = 1$, where $\boldsymbol{\delta}_i$ is the vector of Kronecker deltas $(\delta_{i1}, \dots, \delta_{i|\mathcal{N}|})^{\mathsf{T}}$. This initial condition simply states that the probability of finding a component of one type i node in a graph containing one type i node is 1. Should the graph be of size $\boldsymbol{n}$ instead, this same component would be obtained with a probability equal to the probability that none of the potential directed edges leaving this type i node exist. Considering a generic component $\boldsymbol{l}$, this can be written in mathematical terms as

$$W_i(\boldsymbol{l}|\boldsymbol{n}) = W_i(\boldsymbol{l}|\boldsymbol{l}) \prod_{j \in \mathcal{N}} \binom{n_j - \delta_{ij}}{l_j - \delta_{ij}} \prod_{k \in \mathcal{N}} (1 - p_{jk})^{l_j(n_k - l_k)} \; . \tag{4.1a}$$

In other words, from an initial node of type i, Eq. (4.1a) computes the probability of finding a component of size $\boldsymbol{l}$ in a graph of size $\boldsymbol{n}$ (with $n_j \geq l_j$ for all j) knowing the probability $W_i(\boldsymbol{l}|\boldsymbol{l})$ of finding a component of size $\boldsymbol{l}$ in a graph of size $\boldsymbol{l}$ (i.e., none of the directed edges from these $\boldsymbol{l}$ nodes towards the remaining nodes of the graph must exist). This missing probability is obtained by noting that the distribution $\{W_i(\boldsymbol{m}|\boldsymbol{l})\}$ must be normalized for a given graph size $\boldsymbol{l}$, hence

$$W_i(\boldsymbol{l}|\boldsymbol{l}) = 1 - \sum_{\boldsymbol{m} < \boldsymbol{l}} W_i(\boldsymbol{m}|\boldsymbol{l}) \; , \tag{4.1b}$$

[2] An undirected edge between a type i node and a type j node (noted $i \leftrightarrow j$) therefore occurs with a probability $p_{ij}p_{ji}$. The symmetric case where $p_{ij} = p_{ji}$ for all i and j is statistically equivalent to the situation where all edges are undirected.

where the sum covers every possible instances of $\boldsymbol{m}$ such that $m_j \leq l_j$ for all j but excludes the case where $m_j = l_j$ for each j. Equations. (4.1a)–(4.1b) are mutually dependent: the left-hand side of one feeds the right-hand side of the other. Thus, from a graph consisting of a single node (the initial condition), Eqs. (4.1a)–(4.1b) extend the graph to the desired size $\boldsymbol{n}$, and keep track of the component size distribution along the way to build the final distribution $\{W_i(\boldsymbol{l}|\boldsymbol{n})\}$.

A mass of information is produced during the iteration of Eqs. (4.1a)–(4.1b): the probability of finding every possible components $\boldsymbol{l}$ in every intermediate graph whose size is smaller than or equal to $\boldsymbol{n}$. When interested in bond percolation solely (as in Ref. [7]), the only quantities of interest are the ones related to the graph of maximum size $\boldsymbol{n}$. This ultimately leaves most of the calculated probabilities unused. However, by weighting these unused probabilities with the probability for a graph of intermediate size to occur if nodes were to *exist* with a given probability, the calculation scheme for bond percolation—Eqs. (4.1a)–(4.1b)—can be used to predict the outcomes of site percolation of these multitype random graphs as well.

The probability for a graph of original size $\boldsymbol{n}$ to be left with $\boldsymbol{b}$ nodes after each of its nodes has been independently *kept* with probabilities $\{r_j\}_{j\in\mathcal{N}}$ (i.e., type j nodes are kept with probability r_j) is

$$B_i(\boldsymbol{b}|\boldsymbol{n}) \equiv \prod_{j\in\mathcal{N}} \binom{n_j - \delta_{ij}}{b_j - \delta_{ij}} r_j^{b_j-\delta_{ij}} \bar{r}_j^{n_j-b_j} \ ,$$

where we assume an existing initial node of type i, and note $\bar{r}_j \equiv 1 - r_j$. Hence, from a starting node of type i, the probability $Q_i(\boldsymbol{l}|\boldsymbol{n})$ to find a component of size $\boldsymbol{l}$ in a graph of original size $\boldsymbol{n}$ is

$$Q_i(\boldsymbol{l}|\boldsymbol{n}) = \sum_{\boldsymbol{b}=\boldsymbol{l}}^{\boldsymbol{n}} W_i(\boldsymbol{l}|\boldsymbol{b}) B_i(\boldsymbol{b}|\boldsymbol{n}) = \sum_{\boldsymbol{b}=\boldsymbol{l}}^{\boldsymbol{n}} W_i(\boldsymbol{l}|\boldsymbol{b}) \prod_{j\in\mathcal{N}} \binom{n_j - \delta_{ij}}{b_j - \delta_{ij}} r_j^{b_j-\delta_{ij}} \bar{r}_j^{n_j-b_j} \ , \tag{4.1c}$$

where the two sums cover every possible instances of $\boldsymbol{b}$ such that $l_j \leq b_j \leq n_j$ for every $j \in \mathcal{N}$. Thus we see that by slightly increasing the computational effort, it is possible to incorporate site percolation into the systematic method introduced in Ref. [7] for bond percolation.

4.4.2 Arbitrary graphs

Our framework can also be used to predict the outcomes of hybrid percolation on small arbitrary graphs. By arbitrary graphs, we mean graphs that have a fixed structure in which edges may be directed and/or multiple, and whose nodes may belong to types. We use the *adjacency* matrix $\mathbf{A}$, whose element A_{ij} is the number of edges leaving node i toward node j, to specify the structure. Figure 4.1(a) depicts an example of such an arbitrary graph. Percolation then corresponds to the random removal of edges and nodes according to some given probabilities which may depend on the type of the nodes involved. Predicting the percolation outcomes consists in predicting the probability for a component of size $\boldsymbol{l}$ to be reached from a specific node in a graph originally of size $\boldsymbol{n}$.

For Eqs. (4.1a)–(4.1c) to be applicable, we need to map arbitrary graphs unto multitype random graphs. This mapping is achieved by assigning to each node its own type ($|\mathcal{N}|$ equals the number of nodes), and by setting the probabilities $\{p_{ij}\}$ to mimic the structure of the original arbitrary graph. Consequently, to account for the fact that more than one edge may exist between two nodes in the original graph, we set $p_{ij} = 1 - (1 - \tilde{p}_{ij})^{A_{ij}}$, where $\tilde{p}_{ij}$ is the probability that an edge from node i to node j remains after the random removal of edges in the arbitrary graph. Note that $\tilde{p}_{ij}$ can depend on the type of nodes i and j in the original graph [e.g., there are two types of nodes in Fig. 4.1(a)]. Similarly, the same applies for the existence probabilities of nodes (i.e., $\{r_i\}$ must be equal to the probabilities in the original graph $\{\tilde{r}_i\}$). Using this mapping, Eqs. (4.1a)–(4.1c) offer a systematic procedure to compute the outcomes of hybrid percolation on arbitrary graphs. Figure 4.1(b) compares the predictions of Eqs. (4.1a)–(4.1c) with the results of numerical simulations for the arbitrary graph shown in Fig. 4.1(a). As expected, a perfect agreement is observed.

4.5 Percolation on correlated and clustered infinite random graphs

We now turn our attention to the generalized version of the CM briefly described in the Introduction. We provide a formal definition of the model, and analytically solve percolation for this general ensemble of random graphs. At the end of this section, Table 4.1 summarizes most of the mathematical objects of the formalism.

4.5.1 A *stub matching* scheme

The CM defines an ensemble of graphs that are random in all respects except for the *degree* of their nodes (the number of neighbors) which is prescribed by a given distribution $\{P(k)\}_{k\in\mathbb{N}}$. More precisely, to generate graphs of this ensemble we start with N nodes and assign a degree to each one by drawing an integer from $\{P(k)\}_{k\in\mathbb{N}}$. We then build a list of stubs (half-edges) in which a node whose degree is k appears k times. We shuffle the list and pair nodes in positions $2i$ and $2i+1$ to create edges. Up to corrections of order $\mathcal{O}(N^{-1})$, this procedure uniformly samples the ensemble of graphs with a given degree distribution [144]. Moreover, as closed loops also occurs with a probability proportional to N^{-1}, this procedure generates graphs that are locally tree-like in the limit $N \gg 1$.

We generalize this scheme to account for types of nodes and types of stubs. In our model, each node belongs to a type and we note $\mathcal{N}$ the set of node types, as in the last section. We also note w_i the fraction of nodes whose type is i. As in the CM, nodes are assigned a number of stubs, but now these stubs are identified with types as well. We say that a node has k_α stubs of type α, and we note $\mathcal{E}$ the set of stub types. Unless specified otherwise, greek and latin letters refer to types of edges and nodes, respectively. The number of stubs of each type that a node of type i has is prescribed by the joint degree distribution $\{P_i(k_1, \ldots, k_{|\mathcal{E}|})\}_{k_1,\ldots,k_{|\mathcal{E}|}\in\mathbb{N}} \equiv \{P_i(\boldsymbol{k})\}_{\boldsymbol{k}\in\mathbb{N}^{|\mathcal{E}|}}$. Hence, when generating graphs from this ensemble, each of the N nodes is assigned a type according to $\{w_i\}_{i\in\mathcal{N}}$ and then assigned a number of stubs of each type according to the corresponding joint degree distribution.

To generate graphs from this sequence of nodes, we build a list of stubs for each pair (α, i) where $\alpha \in \mathcal{E}$ and $i \in \mathcal{N}$. For example, a node of type i that has k_α stubs of type α and k_β stubs of type β appears k_α times in the list (α, i) and k_β times in the list (β, i). Stubs are then randomly matched according to a set of rules—noted $\mathcal{R}$—to generate graphs. The information encoded in these rules is twofold. On the one hand, they prescribe in *which* lists should stubs be picked during the matching step. Mathematically, this is encoded in the distribution $\{R(n_{11}, \ldots, n_{|\mathcal{E}|1}, n_{12}, \ldots, n_{|\mathcal{E}||\mathcal{N}|})\}_{n_{11},\ldots,n_{|\mathcal{E}||\mathcal{N}|}\in\mathbb{N}} \equiv \{R(\boldsymbol{n})\}_{\boldsymbol{n}\in\mathbb{N}^{|\mathcal{E}||\mathcal{N}|}}$ giving the proportion of edges—or *hyperedges*, if more than two nodes are involved—that are built by matching $n_{\alpha i}$ stubs of type α stemming from nodes of type i (for every $\alpha \in \mathcal{E}$ and $i \in \mathcal{N}$).

On the other hand, the rules $\mathcal{R}$ prescribe *how* the nodes are connected to one another within the hyperedge. For example, stubs from the list (α, i) and (α, j) could be paired to create undirected edges between layers i and j of *multilayer* graphs. Similarly, stubs from the lists (β, i) and (γ, i) could be paired to create directed edges between nodes of a same type (the two types of stubs corresponding respectively to the *in*-degree and *out*-degree). Moreover, three stubs from a same list could be matched to create triangles, or m stubs of type ε stemming from different types of nodes could be matched to form a multitype Erdős-Rényi motif where edges exist with probability p (see Sec. 4.4.1). In fact, the hyperedges can take any imaginable form and composition as long as they can be handled by Eqs. (4.1). Note that only one stub is required to be part of an hyperedge, even if this hyperedge contributes to more than one to the degree of nodes. For instance, if stubs of type Δ correspond to triangles, a node with $k_\Delta = 2$ will belong to two triangles.

For this graph ensemble to be consistent, the distributions $\{P_i(\boldsymbol{k})\}_{\boldsymbol{k}\in\mathbb{N}^{|\mathcal{E}|}}$ and $\{R(\boldsymbol{n})\}_{\boldsymbol{n}\in\mathbb{N}^{|\mathcal{E}||\mathcal{N}|}}$ must obey certain constraints in the limit $N \gg 1$

$$\frac{w_i \langle k_\alpha \rangle_{P_i}}{w_j \langle k_\nu \rangle_{P_j}} = \frac{\langle n_{\alpha i} \rangle_R}{\langle n_{\nu j} \rangle_R} \tag{4.2}$$

for each $i, j \in \mathcal{N}$ and $\alpha, \nu \in \mathcal{E}$. These constraints simply require that the ratio of the average number of elements in each list (left) equals the relative proportion in which pairs appear in hyperedges (right).

As for the CM, this *stub matching* scheme uniformly samples—up to corrections of order $\mathcal{O}(N^{-1})$—a maximally random ensemble of graphs defined by the distributions $\{w_i\}_{i\in\mathcal{N}}$ and $\{P_i(\boldsymbol{k})\}_{i\in\mathcal{N};\boldsymbol{k}\in\mathbb{N}^{|\mathcal{E}|}}$, and by the rules $\mathcal{R}$. Since stubs are matched randomly, the graphs of that ensemble have a tree-like structure in the limit $N \gg 1$ except within clustered hyperedges.

4.5.2 Probability generating functions

To solve percolation on this general ensemble of random graphs, we adapt the well-known pgf approach [6, 147] to account for node and stub types. As mentioned above, this approach assumes that the structure of the graphs is locally tree-like, an assumption that is not valid whenever an hyperedge contains a loop (e.g., a triangle). However, by solving the component size distribution on each hyperedge beforehand, it is possible to consider that the hyperedge has an *effective* tree-like structure: the probability that there is an *effective* edge from node A to node B is simply the probability that node B can be reached from node A either directly or through the other nodes of the hyperedge. Figures 4.2(a)–(b) illustrate the idea behind the effective tree-like structure. This slight change of perspective allows the use of the pgf approach even though the assumption is not valid in the original graph ensemble.

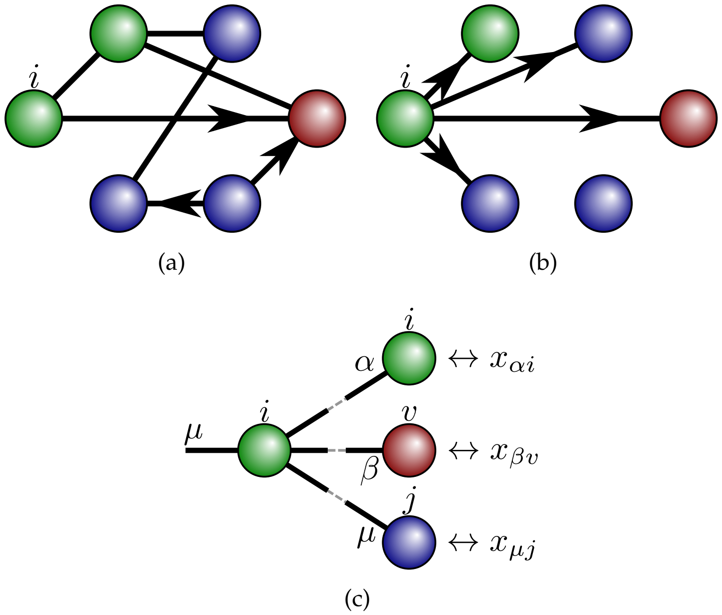
The effective tree-like structure of hyperedges is unveiled with Eqs. (4.1), where a node is now identified by the pair (α, i) instead of by its node type solely. In other words, we keep track of the type of the nodes but also the type of the stubs through which they are involved in the hyperedge. As a result, bold variables like $\boldsymbol{n}$ and $\boldsymbol{l}$ now contain $|\mathcal{E}| \times |\mathcal{N}|$ elements instead of $|\mathcal{N}|$ elements as in Sec. 4.4. The quantity $Q_{\alpha i}(\boldsymbol{l}|\boldsymbol{n};\mathcal{R})$ therefore corresponds to the probability that a pair (α, i) leads to $\boldsymbol{l}$ pairs—i.e., $l_{\nu j}$ pairs (ν, j), for each $\nu \in \mathcal{E}$ and $j \in \mathcal{N}$—in an hyperedge containing $\boldsymbol{n}$ pairs. A dependency on the rules $\mathcal{R}$ has been added in $Q_{\alpha i}(\boldsymbol{l}|\boldsymbol{n};\mathcal{R})$ to explicitly mark that the inner structure of the hyperedges (e.g., quenched or random nature, probabilities of existence of nodes or edges) is prescribed by these rules [3].

The pgf that generates the distribution of the number of pairs that can be reached from an initial pair (α, i) in an hyperedge containing $\boldsymbol{n}$ pairs is

$$\sum_{\boldsymbol{l}=\boldsymbol{\delta}_{\alpha i}}^{\boldsymbol{n}} Q_{\alpha i}(\boldsymbol{l}|\boldsymbol{n};\mathcal{R}) \prod_{\substack{\nu\in\mathcal{E}\\ j\in\mathcal{N}}} x_{\nu j}^{l_{\nu j}-\delta_{\alpha\nu}\delta_{ij}} \tag{4.3}$$

where the sum covers every possible instances of $\boldsymbol{l}$ such that $\delta_{\mu\alpha}\delta_{mi} \leq l_{\mu m} \leq n_{\mu m}$ for every $m \in \mathcal{N}$ and $\mu \in \mathcal{E}$. These two deltas of Kronecker account for the fact that there is at least one pair (α, i) in the hyperedge. Similarly, the two deltas of Kronecker $\delta_{\alpha\nu}\delta_{ij}$ appearing in Eq. (4.3) remove the initial pair—by definition included in $\boldsymbol{l}$—from the count of reachable pairs. Since we are ultimately interested in the number of pairs that can be reached in an hyperedge that has been reached from a given pair, we must remove the dependency of Eqs. (4.3) on the composition $\boldsymbol{n}$ of the hyperedge. To do so, we average this pgf over the probabilities that the

[3] It is not forbidden for two hyperedges to have the same composition $\boldsymbol{n}$ while having different nature or different rules of connection. In such case, the probabilities $Q_{\alpha i}(\boldsymbol{l}|\boldsymbol{n};\mathcal{R})$ should represent the appropriate weighted average of the probabilities computed with Eqs. (4.1) for each hyperedge. It is however strongly encouraged to use different types of stubs—hence different compositions—for each different categories of hyperedges to keep the analysis as clear as possible.


i
i
i
α
$\leftrightarrow x_{\alpha i}$
μ
i
v
$\leftrightarrow x_{\beta v}$
β
j
μ
$\leftrightarrow x_{\mu j}$

(a)

(b)

(c)

FIGURE 4.2 – (a)–(b) Effective tree-like structure of an hyperedge from the point of view of a node of type i. There exists an effective edge between the initial node of type i and any nodes that are directly or indirectly reachable from it. The probability for an effective edge to exist corresponds to the probability that a direct or indirect path exists. (c) Schematic representation of the pgf $f_{\mu i}(\boldsymbol{x})$. Knowing that a node of type i has been reached from one of its stubs of type μ, this pgf generates the distribution of the number of nodes of each type in its neighborhood, as well as the type of stubs from which they have been reached. The types of the nodes and of the stubs are identified with the subscripts of the variables $\boldsymbol{x} = \{x_{\nu i}\}$.

initial pair (α, i) belongs to an hyperedge whose composition is $\boldsymbol{n}$. Because stubs are matched randomly, a random pair (α, i) is ten times more likely to belong to an hyperedge containing ten such pairs than to one that contains only one. Consequently the probabilities $R(\boldsymbol{n})$ must be weighted by the number of pairs (α, i) each composition contains, i.e., averaged over $n_{\alpha i}R(\boldsymbol{n})/\langle n_{\alpha i}\rangle_R$, where the normalizing factor $\langle n_{\alpha i}\rangle_R$ is the average value of $n_{\alpha i}$ with respect to the distribution $R(\boldsymbol{n})$. Doing so yields the pgf generating the distribution of the number of pairs of each types that can be reached from a pair (α, i)

$$\theta_{\alpha i}(\boldsymbol{x}) = \sum_{\boldsymbol{n}} \frac{n_{\alpha i} R(\boldsymbol{n})}{\langle n_{\alpha i}\rangle_R} \sum_{\boldsymbol{l}=\boldsymbol{\delta}_{\alpha i}}^{\boldsymbol{n}} Q_{\alpha i}(\boldsymbol{l}|\boldsymbol{n};\mathcal{R}) \prod_{\substack{\nu\in\mathcal{E} \\ j\in\mathcal{N}}} x_{\nu j}^{l_{\nu j}-\delta_{\alpha\nu}\delta_{ij}} \ , \tag{4.4}$$

where the sum over $\boldsymbol{n}$ covers all hyperedge compositions. Computed for each initial pair (α, i), $\theta_{\alpha i}(\boldsymbol{x})$ provides the projection of the outcomes of percolation on the hyperedges unto an effective branching tree and therefore permits the use of the pgf approach.

To solve percolation on the graphs defined in the previous subsection, we first need to compute the distribution of the composition of the neighborhood of nodes. The neighborhood of a node is the set of *reachable* nodes with which it shares an hyperedge. In other words, node B is a neighbor of node A if there exists an effective edge from node A to node B. The pgf $\theta_{\alpha i}(\boldsymbol{x})$ generates the distribution of neighbors that a node of type i has through one of its stubs of type α. In the limit $N \gg 1$, the tree-like structure of the graphs ensures that the neighboring nodes reachable through two different stubs do not overlap. Hence, the composition of the neighborhood of a node of type i that has k_α and k_β stubs of type α and β is generated by $\left[\theta_{\alpha i}(\boldsymbol{x})\right]^{k_\alpha} \cdot \left[\theta_{\beta i}(\boldsymbol{x})\right]^{k_\beta}$, which corresponds to the convolution of the distributions. Since the number of stubs that nodes of type i have is distributed according to $\{P_i(\boldsymbol{k})\}_{\boldsymbol{k}\in\mathbb{N}^{|\mathcal{E}|}}$, we obtain that the distribution of the composition of the neighborhood of nodes of type i is generated by

$$g_i(\boldsymbol{x}) = \sum_{\boldsymbol{k}} P_i(\boldsymbol{k}) \prod_{\alpha\in\mathcal{E}} \left[\theta_{\alpha i}(\boldsymbol{x})\right]^{k_\alpha} , \tag{4.5}$$

where the sum covers all cases where $P_i(\boldsymbol{k}) \neq 0$. As in $\theta_{\alpha i}(\boldsymbol{x})$, this pgf keeps track of the type of the stubs from which the neighboring nodes have been reached through the subscripts of the variables $\boldsymbol{x} = \{x_{\nu j}\}_{\nu\in\mathcal{E};j\in\mathcal{N}}$. In other words, $g_i(\boldsymbol{x})$ generates the number of pairs that are in the neighborhood of a node of type i. This pgf is analogous to the function $G_0(x)$ generating the degree distribution in the CM [144, 147].

To complete the solution of the percolation problem requires the knowledge of the distribution of the composition of the neighborhood of nodes that have been reached from one of their stubs. As discussed for $\theta_{\alpha i}(\boldsymbol{x})$, the probability for a stub of type μ to be attached to a node with a total of $\boldsymbol{k}$ stubs (i.e., k_α stubs of type α for every $\alpha \in \mathcal{E}$) is weighted by the number of stubs of type μ that this node has. Hence, given that a node of type i has been reached through one of its stubs of type μ [i.e., a pair (μ, i)], the composition of the neighborhood accessible from its *other* stubs is generated by

$$f_{\mu i}(\boldsymbol{x}) = \sum_{\boldsymbol{k}} \frac{k_\mu P_i(\boldsymbol{k})}{\langle k_\mu \rangle_{P_i}} \prod_{\alpha\in\mathcal{E}} \left[\theta_{\alpha i}(\boldsymbol{x})\right]^{k_\alpha - \delta_{\alpha\mu}} , \tag{4.6}$$

where the delta $\delta_{\mu\nu}$ has been added to exclude from the count the stub of type μ from which the node has been reached, and where $\langle k_\mu \rangle_{P_i}$ is the average number of stubs of type μ that nodes of type i have. The distributions generated by $f_{\mu i}(\boldsymbol{x})$ are analogous to the *excess* degree distribution generated by $G_1(x)$ in the CM [144, 147]. Figure 4.2(c) illustrates the information encoded in the pgfs $f_{\mu i}(\boldsymbol{x})$.

4.5.3 Extensive "giant" component

Having defined the pgfs $g_i(\boldsymbol{x})$ and $f_{\mu i}(\boldsymbol{x})$, the behavior of the extensive "giant" component can be predicted in the limit $N \gg 1$ using simple self-consistency arguments. We define $a_{\mu i}$ as the probability that a node of type i reached via one of its stubs of type μ does *not* lead to the giant component. Self-consistency then requires that if this pair does not lead to the giant component, then neither should the pairs that are reachable from it. Since the distribution of the number of pairs reachable from a given pair (μ, i) is generated by Eq. (4.6), this self-consistency requirement can be rewritten as

$$a_{\mu i} = f_{\mu i}(\boldsymbol{a}) \tag{4.7}$$

for every $\mu \in \mathcal{E}$ and $i \in \mathcal{N}$. Because the coefficients of $f_{\mu i}(\boldsymbol{x})$ are normalized (they form a probability distribution), the point $\boldsymbol{a} = \mathbf{1}$ (every $a_{\mu i}$ equals 1) is always a solution of Eqs. (4.7). However, as the density of edges and/or nodes increases, another solution where at least one element of $\boldsymbol{a}$ is smaller than 1 appears. This new solution marks the emergence of an extensive component.

Because their coefficients are all positives, the pgfs $f_{\mu i}(\boldsymbol{x})$ are all convex and monotonic increasing in $[0,1]^{|\mathcal{E}||\mathcal{N}|}$. Hence when $\boldsymbol{a} = \mathbf{1}$ is the only solution of Eqs. (4.7) in $[0,1]^{|\mathcal{E}||\mathcal{N}|}$, it is the stable fixed point of (with $n \in \mathbb{N}$)

$$\boldsymbol{a}^{(n+1)} = \boldsymbol{f}\big(\boldsymbol{a}^{(n)}\big) \ , \tag{4.8}$$

for any initial condition $\boldsymbol{a}^{(0)}$ in $[0,1]^{|\mathcal{E}||\mathcal{N}|}$, and where the map $\boldsymbol{f}(\boldsymbol{x})$ consists of every $f_{\mu i}(\boldsymbol{x})$. This fixed point becomes unstable through a transcritical bifurcation as soon as another solution in $[0,1]^{|\mathcal{E}||\mathcal{N}|}$ appears. In fact, the shape of $\boldsymbol{f}(\boldsymbol{x})$ in $[0,1]^{|\mathcal{E}||\mathcal{N}|}$ and the fact that $\boldsymbol{f}(\mathbf{1}) = \mathbf{1}$ implies that this other solution is unique in the interval of interest, that it is a stable fixed point of Eq. (4.8), and that the transition is continuous. Analyzing the stability of $\boldsymbol{f}(\boldsymbol{x})$ around the fixed point $\boldsymbol{a} = \mathbf{1}$ leads to the criterion for the emergence of the giant component

$$\det(\mathbf{J} - \mathbf{I}) = 0 \ , \tag{4.9}$$

where $\mathbf{J}$ is the Jacobian matrix of $\boldsymbol{f}(\boldsymbol{x})$ around $\boldsymbol{x} = \mathbf{1}$, and $\mathbf{I}$ is the identity matrix. Put differently, an extensive component exists whenever the largest eigenvalue of $\mathbf{J}$, $\lambda_{\max}(\mathbf{J})$, is greater than one [4].

Having solved Eqs. (4.7), the probability that a node of type i leads to the giant component through at least one of its neighbors is given by $\mathcal{P}_i = 1 - g_i(\boldsymbol{a})$. Consequently, the probability that a randomly chosen existing node does lead to the giant component is

$$\mathcal{P} = \sum_{i \in \mathcal{N}} \frac{r_i w_i \mathcal{P}_i}{\sum_{j \in \mathcal{N}} r_j w_j} = 1 - \sum_{i \in \mathcal{N}} \frac{r_i w_i g_i(\boldsymbol{a})}{\sum_{j \in \mathcal{N}} r_j w_j} \ , \tag{4.10}$$

where r_i is the probability that a node of type i exists.

As shown in Sec. 4.4, hyperedges may include directed edges or edges that are more likely to exist in one direction than the other [i.e., $p_{ij} \neq p_{ji}$ in Eqs. (4.1a)]. This implies that while node B is in the neighborhood of node A, node A may not be in the neighborhood of node B. From

[4] Because the coefficients of $\boldsymbol{f}(\boldsymbol{x})$ are all positives, $\mathbf{J}$ is a non-negative matrix and the Perron-Frobenius theorem ensures that its largest eigenvalue is real, positive and non-degenerate.

such local asymmetries, a global asymmetry arises between the probability that a node leads to the giant component ($\mathcal{P}$) and the relative size $\mathcal{S}$ of the giant component. In such case, the extensive component has a "bow-tie" structure [147, 10] meaning that the nodes involved in the extensive component belong to one of the three non-overlapping sets $\mathcal{I}^{\text{in}}$, $\mathcal{I}^{\text{both}}$ and $\mathcal{I}^{\text{out}}$. The set $\mathcal{I}^{\text{in}}$ includes nodes that lead to the giant component but that cannot be reached from it; these nodes are somehow "hidden" behind directed edges. The set $\mathcal{I}^{\text{out}}$ contains nodes that cannot lead to the giant component but that can be reached from it; they are positioned downstream of directed edges. The set $\mathcal{I}^{\text{both}}$ contains nodes that lead to the giant component and that can be reached from it. From this, we conclude $\mathcal{P} = |\mathcal{I}^{\text{in}} \bigcup \mathcal{I}^{\text{both}}|/N$ and $\mathcal{S} = |\mathcal{I}^{\text{both}} \bigcup \mathcal{I}^{\text{out}}|/N$.

This perspective offers a direct and intuitive way to calculate $\mathcal{S}$: it is the probability that a node does not lead to the extensive component when the direction of every edges is reversed. This edge reversal is fully encoded in $\bar{Q}_{\alpha i}(\boldsymbol{l}|\boldsymbol{n};\mathcal{R})$ computed with Eqs. (4.1) with incoming directed edges swapped into outgoing ones (and vice versa), and with edges that were more likely to exist in a given direction now more likely to exist in the opposite direction (i.e., p_{ij} becomes p_{ji}). From these probabilities, we define the pgfs $\bar{\theta}_{\alpha i}(\boldsymbol{x})$, $\bar{g}_i(\boldsymbol{x})$ and $\bar{f}_{\mu i}(\boldsymbol{x})$ which are analogous to the ones previously defined [$P_i(\boldsymbol{k})$ and $R(\boldsymbol{n})$ remain unchanged]. Defining $\bar{a}_{\mu i}$ as the probability that a node of type i reached by one of its stubs of type μ does not lead to the giant component in the *reversed* graph ensemble, self-consistency now requires

$$\bar{a}_{\mu i} = \bar{f}_{\mu i}(\bar{\boldsymbol{a}}) \tag{4.11}$$

for every $\mu \in \mathcal{E}$ and $i \in \mathcal{N}$. As for Eqs. (4.7), the solution of this set of equations correspond to the fixed point of the corresponding map and can therefore be obtained by successive iterations of any initial condition in $[0,1]^{|\mathcal{E}||\mathcal{N}|}$. The elements of the Jacobian matrix of both Eqs. (4.7) and (4.11) are the average number of pairs, say (α, j), that are in the neighborhood of a pair, say (μ, i), in their respective graph ensemble. Since both systems, Eqs. (4.7) and (4.11), correspond to different perspectives of the same graph ensemble, the two Jacobian matrices are mathematically similar, and therefore have the same eigenvalues. Hence the transcritical bifurcation occurs simultaneously in both systems.

Having obtained $\bar{\boldsymbol{a}}$ from Eqs. (4.11), the probability for a node of type i to be part of the giant component is $\mathcal{S}_i = 1 - \bar{g}_i(\bar{\boldsymbol{a}})$, and the relative size of the giant component is

$$\mathcal{S} = \sum_{i\in\mathcal{N}} \frac{r_i w_i \mathcal{S}_i}{\sum_{j\in\mathcal{N}} r_j w_j} = 1 - \sum_{i\in\mathcal{N}} \frac{r_i w_i \bar{g}_i(\bar{\boldsymbol{a}})}{\sum_{j\in\mathcal{N}} r_j w_j} \, . \tag{4.12}$$

Clearly, when all hyperedges are symmetric (i.e., $p_{ij} = p_{ji}$ for every $i, j \in \mathcal{N}$) there is no global asymmetry in the graph ensemble, and $\mathcal{P} = \mathcal{S}$. Also, whenever Eqs. (4.3)–(4.12) are used in the context of site percolation—where nodes exist or are activated with a given set of probabilities—the value of $\mathcal{P}$ and $\mathcal{S}$ is relative to the number of nodes that exist. In other words, $\mathcal{P}$ is the probability that an existing node leads to an extensive component, and $\mathcal{S}$ is the fraction of existing nodes that are part of it.

4.5.4 Small components

Substituting $x_{\nu j}$ by z_j for every $j \in \mathcal{N}$ and $\nu \in \mathcal{E}$ in Eq. (4.5) yields a pgf that generates the number of nodes of each type that are *directly accessible* from a node of type i (i.e., nodes that are in its neighborhood). Using self-consistency arguments similar to the one used in the previous subsection, it is possible to obtain a pgf that generates the distribution of the number of nodes of each type that will be *eventually reached* from a node of type i; the *reach* of this new pgf is no longer limited to the immediate neighborhood. In fact this new pgf allows to investigate the composition and the sizes of the components that contain a finite number of nodes. Let this new pgf be denoted $K(\boldsymbol{z})$.

To compute $K(\boldsymbol{z})$, we first consider the pgf $A_{\alpha i}(\boldsymbol{x})$ that generates the distribution of the number of nodes of each type that can be reached from a node of type i given that this node has been reached from one of its stubs of type α. Note that $A_{\alpha i}(\boldsymbol{x})$ is a function of $\boldsymbol{x}$ so that it keeps track of the type of the stubs from which each node has been reached. Besides yielding a tree-like structure, the *stub matching* scheme used to generate graphs implies that the pgfs $\{f_{\mu i}(\boldsymbol{x})\}$ are invariant under translations on the graphs in the limit $N \gg 1$. In other words, while navigating on a graph from this ensemble, the number and the type of the nodes downstream from any given node does not depend on the types of the nodes (or the types of the stubs) previously encountered; navigating on graphs from this ensemble is a stationary Markov process (i.e., it only depends on the *current position* on the graph). Consequently, a node of type i reached from one of its stubs of type α and a pair (α, i) present in its neighborhood should both lead to a finite tree whose size and composition are identically distributed; this distribution is generated by $A_{\alpha i}(\boldsymbol{x})$. Considering every combination (α, i), this self-consistency requirement can be mathematically formulated as

$$A_{\alpha i}(\boldsymbol{x}) = x_{\alpha i} f_{\alpha i}\big(\boldsymbol{A}(\boldsymbol{x})\big) \ , \tag{4.13}$$

where the extra $x_{\alpha i}$ accounts for the node of type i that has been reached through one of its stubs of type α. Analogously to the set of probabilities $\{a_{\alpha i}\}$, the pgfs $\{A_{\alpha i}(\boldsymbol{x})\}$ are the fixed point of (with $n \in \mathbb{N}$)

$$\boldsymbol{A}^{(n+1)}(\boldsymbol{x}) = \boldsymbol{x} \circ \boldsymbol{f}\big(\boldsymbol{A}^{(n)}(\boldsymbol{x})\big) \tag{4.14}$$

where "$\circ$" denotes the elementwise multiplication. It is in fact straightforward to show that the extra $\boldsymbol{x}$ guarantees that the distributions generated by $\boldsymbol{A}(\boldsymbol{x})$ can be obtained for components of n nodes or less in $n+1$ iterations of Eq. (4.14) from the initial condition $\boldsymbol{A}^{(0)}(\boldsymbol{x}) = \boldsymbol{1}$.

Having obtained $\boldsymbol{A}(\boldsymbol{x})$ up to a sufficient component size n, the number of nodes of each type that can be reached in a finite component from a randomly chosen node of type i is generated by $K_i(\boldsymbol{z}) \equiv z_i g_i\big(\boldsymbol{A}(\boldsymbol{z})\big)$. The pgf generating the number of nodes of each type that are accessible in a small component from a randomly chosen existing node is

$$K(\boldsymbol{z}) = \sum_{i\in\mathcal{N}} \frac{r_i w_i K_i(\boldsymbol{z})}{\sum_{j\in\mathcal{N}} r_j w_j} = \sum_{i\in\mathcal{N}} \frac{r_i w_i z_i g_i\big(\boldsymbol{A}(\boldsymbol{z})\big)}{\sum_{j\in\mathcal{N}} r_j w_j} \ . \tag{4.15}$$

It is worth mentioning that the distributions generated by $K(\boldsymbol{z})$ and $\{A_{\alpha i}(\boldsymbol{z})\}$ are not normalized in the presence of an extensive component as there is a non-zero probability that a pair (α, i) leads to the giant component. In fact, comparing Eqs. (4.8) and (4.14) leads to the conclusion that $A_{\alpha i}(\mathbf{1}) = a_{\alpha i}$ and that $K(\mathbf{1}) = 1 - \mathcal{P}$.

4.5.5 Discontinuous phase transition

There has been a sustained interest in recent years for interdependent—or multiplex—graphs as models of critical interdependent infrastructures like power grids [33, 76, 195]. From the theoretical perspective, the most striking feature of interdependent graphs is the sudden emergence of the extensive component through a discontinuous, or first-order, transition. This indeed contrasts greatly with the continuous, or second order, transition typically encountered in percolation. As shown in Ref. [181], this change in the order of the phase transition can be understood through a change in the definition of what constitutes an extensive component (i.e., the order parameter). We briefly explain and illustrate in this subsection how our approach can also be used to model interdependent graphs.

Interdependent graphs are composed of two or more graphs in which the state of a node depends on the state of its twin nodes in the other graphs. For instance, power stations in power grids strongly rely on the Internet communication network, which in turn depends on the power grid [34]. In these graphs, an active node is considered functional if and only if all of its twin nodes are active as well. Studying percolation on such coupled graphs shifts its focus from the emergence of an extensive component composed of active nodes to the emergence of an extensive component composed of functional nodes. This requires a redefinition of Eqs. (4.7)–(4.12).

To lighten the description, we consider a case of two interdependent graphs, A and B, that are partially dependent—only a fraction of the nodes in A depends on a node in B, and vice versa—and in which all edges are undirected (i.e., no global asymmetry hence $\mathcal{S} = \mathcal{P}$). The pgfs $g_i(\boldsymbol{x})$ and $f_{\mu i}(\boldsymbol{x})$ and all other related quantities are known for both graphs and are identified with the superscript A or B. A node of type i in graph A (type j in graph B) depends on one twin node of type v in graph B (type u in graph A) with probability q_{iv}^{AB} (q_{ju}^{BA}). Since the two graphs are only partially dependent, these probabilities need not to sum to one as some nodes may not depend on a twin node (e.g., $\sum_{v\in\mathcal{N}^B} q_{iv}^{AB} \leq 1$). Twin nodes are chosen at random, and the dependence relationship is not typically reciprocal.

Quantity	Description	Definition
$\mathcal{N}$	Set of the types of nodes	Sec. 4.4.1
$\mathcal{E}$	Set of the types of stubs	Sec. 4.5.1
$\mathcal{R}$	Rules dictating the way stubs are matched to form hyperedges	Sec. 4.5.1
r_i	Probability that a node of type i exists in a multitype random graph	Sec. 4.4.1
p_{ij}	Probability that a directed edge exists from a node of type i towards a node of type j in a multitype random graph	Sec. 4.4.1
w_i	Fraction of nodes that are of type i	Sec. 4.5.1
$P_i(\boldsymbol{k})$	Fraction of nodes of type i that have $\boldsymbol{k}$ stubs	Sec. 4.5.1
$\langle b \rangle_B$	Average value of the quantity b over the distribution $B(b)$	Eq. (4.2)
$R(\boldsymbol{n})$	Distribution of the composition of hyperedges	Sec. 4.5.1
q_{iv}^{AB}	Probability that a node of type i in graph A depends on a node of type v in graph B	Sec. 4.5.5
$W_i(\boldsymbol{l}\|\boldsymbol{n})$	Probability that a node of type i leads to a component of $\boldsymbol{l}$ nodes in a multitype random graph of $\boldsymbol{n}$ nodes where edges exist with probabilities $\{p_{ij}\}$	Eqs. (4.1a)–(4.1b)
$Q_i(\boldsymbol{l}\|\boldsymbol{n})$	Probability that a node of type i leads to a component of $\boldsymbol{l}$ nodes in a multitype random graph of $\boldsymbol{n}$ nodes where nodes and edges exist with probabilities $\{r_i\}$ and $\{p_{ij}\}$ respectively	Eq. (4.1c)
$a_{\mu i}$	Probability that a pair of type (μ, i) does not lead to the giant component	Eq. (4.7)
$\mathcal{P}$	Probability that a random node leads to the giant component	Eq. (4.10)
$\bar{a}_{\mu i}$	Probability that a pair of type (μ, i) cannot be reached from the giant component	Eq. (4.11)
$\mathcal{S}$	Size of the giant component	Eq. (4.12)
$\theta_{\alpha i}(\boldsymbol{x})$	pgf generating the distribution of the number of pairs of each type that can be reached from a pair (α, i)	Eq. (4.4)
$g_i(\boldsymbol{x})$	pgf generating the distribution of the number of pairs of each type that are in the neighborhood of a node of type i	Eq. (4.5)
$f_{\mu i}(\boldsymbol{x})$	pgf generating the distribution of the number of pairs of each type that are in the neighborhood of a node of type i that has been reached by one of its stubs of type μ (the pair (μ, i) accessible from that stub is excluded)	Eq. (4.6)
$A_{\alpha i}(\boldsymbol{x})$	pgf generating the distribution of the composition of the nodes downstream from a pair of type (α, i)	Eq. (4.13)
$K(\boldsymbol{z})$	pgf generating the distribution of the composition of the small component accessible from a random node	Eq. (4.15)

TABLE 4.1 – Glossary of the sets, probabilities, distributions and pgf defined in our approach.

	semi-directed (Sec. 4.6.2)	correlated (Sec. 4.6.3)		clustered (Sec. 4.6.4)
		multitype	degree	edge-triangle
$\mathcal{N}$	$\{0\}$	$\{0,1,2,3,\ldots\}$	$\{0,1,2,3,\ldots,k_{\max}\}$	$\{0\}$
$\mathcal{E}$	$\{A,B,C\}$	$=\mathcal{N}$	$=\mathcal{N}$	$\{A,B\}$
$P_i(\boldsymbol{k})$	$P_0(k_A,k_B,k_C)$	$P_i(k_0,k_1,k_2,\ldots)$	$\frac{i!}{\prod_{j'\in\mathcal{N}} k_{j'}!}\prod_{j\in\mathcal{N}}[P(j\vert i)]^{k_j}$	$P_0(k_A,k_B)$
$\{w_i\}$	$w_0 = 1.0$	fixed with Eq. (4.2)	fixed with Eq. (4.2)	$w_0 = 1.0$
$R(\boldsymbol{n})$	$R(n_{0A},n_{0B},n_{0C})$ $R(2,0,0) = \frac{\langle k_A\rangle_{P_0}}{(\langle k_A\rangle_{P_0}+2\langle k_B\rangle_{P_0})}$ $R(0,1,1) = \frac{2\langle k_B\rangle_{P_0}}{(\langle k_A\rangle_{P_0}+2\langle k_B\rangle_{P_0})}$	$\frac{1}{R'}\sum_{i,j\in\mathcal{N}}(1-\delta_{0,n_{ji}})w_i\langle k_j\rangle_{P_i}$	$\frac{1}{R'}\sum_{i,j\in\mathcal{N}}(1-\delta_{0,n_{ji}})w_i\langle k_j\rangle_{P_i}$	$R(n_{A0},n_{0B})$ $R(2,0) = 3\langle k_A\rangle_{P_0}/(3\langle k_A\rangle_{P_0}+2\langle k_B\rangle_{P_0})$ $R(0,3) = 2\langle k_B\rangle_{P_0}/(3\langle k_A\rangle_{P_0}+2\langle k_B\rangle_{P_0})$
$\theta_{\alpha i}(\boldsymbol{x})$	$\theta_{A0}(\boldsymbol{x}) = (1-rT)+rTx_{A0}$ $\theta_{B0}(\boldsymbol{x}) = (1-rT)+rTx_{C0}$ $\theta_{C0}(\boldsymbol{x}) = 1$	$\theta_{ji}(\boldsymbol{x}) = (1-r_jT_{ij})+r_jT_{ij}x_{ij}$	$\theta_{ji}(\boldsymbol{x}) = (1-r_jT_{ij})+r_jT_{ij}x_{ij}$	$\theta_{A0} = (1-rT)+rTx_{A0}\theta_{B0}$ $\theta_{B0} = (1-rT)^2+2rT[1-rT(2-T)]x_{B0}$ $+r^2T^2[3-2T]x_{B0}^2$
$\bar{\theta}_{\alpha i}(\boldsymbol{x})$	$\bar{\theta}_{A0}(\boldsymbol{x}) = (1-rT)+rTx_{A0}$ $\bar{\theta}_{B0}(\boldsymbol{x}) = 1$ $\bar{\theta}_{C0}(\boldsymbol{x}) = (1-rT)+rTx_{B0}$	$\bar{\theta}_{ji}(\boldsymbol{x}) = (1-r_jT_{ji})+r_jT_{ji}x_{ij}$	$\bar{\theta}_{ji}(\boldsymbol{x}) = (1-r_jT_{ji})+r_jT_{ji}x_{ij}$	$\bar{\theta}_{A0} = (1-rT)+rTx_{A0}\theta_{B0}$ $\bar{\theta}_{B0} = (1-rT)^2+2rT[1-rT(2-T)]x_{B0}$ $+r^2T^2[3-2T]x_{B0}^2$

TABLE 4.2 – Special cases of our formalism that correspond to other widely-used approaches. More details can be found in the corresponding subsections of the main text.

To compute the size of the extensive functional component, we define $a^A_{\mu i}$ as the probability that a node of type i reached from one of its stubs of type μ in graph A does *not* lead to the *functional* extensive component. Due to the tree-like structure of the graphs, the probability that this node does lead to an extensive component in graph A is $[1 - f^A_{\mu i}(\boldsymbol{a^A})]$. Additionally, for this node to belong to the functional component, its twin node must also belong to the giant component in graph B. Since twin nodes are chosen at random, this occurs with probability $[1 - \sum_{v\in\mathcal{N}^B} q^{AB}_{iv} r^B_v g^B_v(\boldsymbol{a^B})]$, where r^B_v is the probability that a node of type v in graph B exists. The same reasoning applied to the twin probability $a^B_{\nu j}$, we find that both sets of probabilities are the solution of

$$a^A_{\mu i} = 1 - \Big[1 - f^A_{\mu i}(\boldsymbol{a^A})\Big]\Big[1 - \sum_{v\in\mathcal{N}^B} q^{AB}_{iv} r^B_v g^B_v(\boldsymbol{a^B})\Big] \tag{4.16a}$$

$$a^B_{\nu j} = 1 - \Big[1 - f^B_{\nu j}(\boldsymbol{a^B})\Big]\Big[1 - \sum_{u\in\mathcal{N}^A} q^{BA}_{ju} r^A_u g^A_u(\boldsymbol{a^A})\Big] \tag{4.16b}$$

for every $i \in \mathcal{N}^A$, $j \in \mathcal{N}^B$, $\mu \in \mathcal{E}^A$ and $\nu \in \mathcal{E}^B$. These equations solved, the probability that a node of type i in graph A (type j in graph B) belongs to the functional component is

$$\mathcal{S}^A_i = \Big[1 - g^A_i(\boldsymbol{a^A})\Big]\Big[1 - \sum_{v\in\mathcal{N}^B} q^{AB}_{iv} r^B_v g^B_v(\boldsymbol{a^B})\Big] \tag{4.17a}$$

$$\mathcal{S}^B_j = \Big[1 - g^B_j(\boldsymbol{a^B})\Big]\Big[1 - \sum_{u\in\mathcal{N}^A} q^{BA}_{ju} r^A_u g^A_u(\boldsymbol{a^A})\Big] \; , \tag{4.17b}$$

which is similar to the calculation of $\mathcal{S}_i$ in Sec. 4.5.3 but where the probability that the node of type i belongs to the functional extensive component is weighted by the probability that its twin node belongs to the extensive functional component as well. Averaging over the fraction of existing nodes of each type in the corresponding graph, we get the fraction of nodes that belong to the functional component in each graph

$$\mathcal{S}^A = \sum_{i\in\mathcal{N}^A} \frac{r^A_i w^A_i}{\sum_{j\in\mathcal{N}^A} r^A_j w^A_j} \mathcal{S}^A_i \tag{4.18a}$$

$$\mathcal{S}^B = \sum_{j\in\mathcal{N}^B} \frac{r^B_i w^B_i}{\sum_{j\in\mathcal{N}^B} r^B_j w^B_j} \mathcal{S}^B_j \; . \tag{4.18b}$$

As for the quantities $\mathcal{P}$ and $\mathcal{S}$ defined previously, the fractions $\mathcal{S}^A$ and $\mathcal{S}^B$ are relative to the number of *existing* nodes. However, contrariwise to Eqs. (4.7) and Eqs. (4.11), the system of equations (4.16) is not composed of monotonously increasing functions (some coefficients in the polynomials are negative). This implies that, although the point $\boldsymbol{a^A} \oplus \boldsymbol{a^B} = \boldsymbol{1}$ is still a solution, an other solution in $[0,1]^{|\mathcal{N}^A||\mathcal{E}^A| + |\mathcal{N}^B||\mathcal{E}^B|}$ corresponding to the presence of an extensive functional component may not appear continuously from $\boldsymbol{a^A} \oplus \boldsymbol{a^B} = \boldsymbol{1}$. Hence the values of $\mathcal{S}^A$ and $\mathcal{S}^B$ *jump* abruptly from zero to a finite value in [0,1] which corresponds to a discontinuous phase transition. As a final remark, Eqs. (4.16)–(4.18) are straightforward to generalize to an arbitrary number of interdependent graphs (see Ref. [181] for guidelines).

A case study involving interdependent graphs is investigated in Sec. 4.6.7. The example illustrates the versatility of our approach as well as its role as a theoretical laboratory to investigate the impact of local properties on the global connectivity of complex graphs.

4.6 Examples and applications

To demonstrate the versatility and the flexibility of the formalism, we present a series of representative examples: this will also clarify the conceptual and numerical steps necessary to implement such a general approach.

4.6.1 Combinatorial interpretation

When carried out analytically, Eqs. (4.1a)–(4.1c) yield polynomials whose coefficients have a direct combinatorial interpretation. Let us introduce the identity for a graph of size $\boldsymbol{l}$

$$1 = \prod_{j,k \in \mathcal{N}} (p_{jk} + \bar{p}_{jk})^{l_j(l_k - \delta_{jk})} \tag{4.19}$$

where $\bar{p}_{jk} \equiv 1 - p_{jk}$. Each term on the right-hand side of Eq. (4.19) corresponds to the probability of a specific configuration of a multitype random graph of size $\boldsymbol{l}$ whose edges exist independently with probabilities $\{p_{jk}\}$. To understand this, let us recall that an edge from a node of type j to a node of type k (noted $j \to k$) exists with probability p_{jk} and does not exist with probability $\bar{p}_{jk}$. If there are l_j and l_k nodes of type j and k respectively, then there are a total of $l_j l_k$ potential $j \to k$ edges if $j \neq k$, or $l_j(l_j - 1)$ if $j = k$. Hence $(p_{jk} + \bar{p}_{jk})^{l_j(l_k - \delta_{jk})}$ is a polynomial in p_{jk} and $\bar{p}_{jk}$ whose elements are of the form $b p_{jk}^{c} \bar{p}_{jk}^{\bar{c}}$, where b is the number of ways a configuration in which a number c of $j \to k$ edges exist and $\bar{c}$ do not exist. Since edges exist independently of one another, the product on the right-hand side of Eq. (4.19) generates every possible configuration of a multitype random graph. Using this identity in Eq. (4.1b) yields a polynomial, $Q_i(\boldsymbol{l}|\boldsymbol{n})$, in which the coefficient in front of the term $\prod_{sjk} r_s^{a_s} \bar{r}_s^{\bar{a}_s} p_{jk}^{c_{jk}} \bar{p}_{jk}^{\bar{c}_{jk}}$ is the number of distinct ways to reach a component of size $\boldsymbol{l}$, from an initial type i node, in a graph of original size $\boldsymbol{n}$ in which a_s type s nodes exist (and $\bar{a}_s$ that do not exist) and where a number b_{jk} of $j \to k$ edges exist (and $\bar{b}_{jk}$ do not exist). Hence, Eqs. (4.1a)–(4.1c) offer a systematic method to enumerate the number of graphs with a given number of edges and labeled nodes of different types.

The polynomials obtained from Eqs. (4.1a)–(4.1c) may be used in the context of "traditional" percolation theory: the study of the formation of clusters on infinite lattices. In fact, several methods such as the *renormalization-group* transformation [164, 165] or the *triangle-triangle* transformation [172, 173] rely on the connectivity of a basic cell to determine or approximate percolation thresholds and critical exponents. These two methods are based on the idea that the percolation of the whole lattice is related to the "percolation" of the basic cells. For renormalization method, the transformation is the polynomial corresponding to the probability that one can “get across” (based on specific criteria) the basic cell. For the triangle-triangle transformation, one of the polynomials involved to compute the percolation threshold is the probability for the three nodes at the vertices of the triangular-shaped basic cell to be somehow reachable from one another. For both of these methods, the required polynomials are simply a combination of the polynomials obtained by substituting Eq. (4.19) into Eqs. (4.1a)–(4.1c). While the calculation of the required polynomials may be easily carried out analytically by hand for simple cells, it becomes rapidly tedious as the size of the basic cell increases. Our method therefore offers a systematic way to compute these polynomials for larger and more complicated basic cells.

4.6.2 Semi-directed random graphs

Semi-directed random graphs are composed of indistinguishable nodes connected via undirected and directed edges. They were used in Ref. [125] to study the impact of non-reciprocal connections in contact networks on the propagation of an emerging infectious disease. These non-reciprocal connections accounted for the susceptibility of health-care workers to get infected from infectious individuals seeking treatments in hospitals. Semi-directed are also a good example as they have the well-known undirected graphs and directed graphs as special cases.

Every nodes in these graphs belong to the same type of node ($|\mathcal{N}| = 1$, type 0, w_0=1), and there are $|\mathcal{E}| = 3$ types of stubs: stubs of type A are paired together to form undirected edges, and stubs of type B and C are paired to form a directed edge from the B stub to the C stub. The joint degree distribution $P_0(\boldsymbol{k}) = P_0(k_A, k_B, k_C)$ corresponds to the distribution of *undirected* degree, *out*-degree and *in*-degree. In this scenario, the conditions given by Eq. (4.2) imply that there must be as much incoming stubs as there are outgoing stubs, $\langle k_B \rangle_{P_0} = \langle k_C \rangle_{P_0}$, and they fix the values of $R(\boldsymbol{n}) = R(n_{0A}, n_{0B}, n_{0C})$ in terms of the average degrees, $R(2,0,0) = 1 - R(0,1,1) = \langle k_A \rangle_{P_0} / (\langle k_A \rangle_{P_0} + 2\langle k_B \rangle_{P_0})$. Assuming that edges exist with probability T and nodes exist with probability r, we find from Eqs. (4.1) and (4.4)

$$\theta_{A0}(\boldsymbol{x}) = (1 - rT) + rTx_{A0} \tag{4.20a}$$

$$\theta_{B0}(\boldsymbol{x}) = (1 - rT) + rTx_{C0} \tag{4.20b}$$

$$\theta_{C0}(\boldsymbol{x}) = 1 \ , \tag{4.20c}$$

from which we define the pgfs $g_0(\boldsymbol{x})$, $f_{A0}(\boldsymbol{x})$ and $f_{C0}(\boldsymbol{x})$ from Eqs. (4.5) and (4.6). Note that $f_{B0}(\boldsymbol{x})$ does not exist as nodes cannot be reached by an outgoing stub. Similarly, when reversing the direction of edges (directed edges now run from C stubs to B stubs), we obtain

$$\bar{\theta}_{A0}(\boldsymbol{x}) = (1 - rT) + rTx_{A0} \tag{4.21a}$$

$$\bar{\theta}_{B0}(\boldsymbol{x}) = 1 \tag{4.21b}$$

$$\bar{\theta}_{C0}(\boldsymbol{x}) = (1 - rT) + rTx_{B0} \ , \tag{4.21c}$$

which yield the pgfs $\bar{g}_0(\boldsymbol{x})$, $\bar{f}_{A0}(\boldsymbol{x})$ and $\bar{f}_{B0}(\boldsymbol{x})$ [$\bar{f}_{C0}(\boldsymbol{x})$ is ill-defined]. Using Eqs. (4.20) and (4.21) in Eqs. (4.5)–(4.15) with $r = 1$ yields the results obtained in Ref. [125], and to the ones obtained for purely directed [147] or purely undirected random graphs [136] in the appropriate limits. A summary of this subsection as well as of the following two subsections is given in Table 4.2.

4.6.3 Correlated random graphs

Other interesting special cases of our model are correlated random graphs: graphs where nodes are more likely to be connected with nodes having specific intrinsic properties (e.g., degree, centrality, ethnicity, age group, gender). In such cases, there are $|\mathcal{N}|$ types of nodes, one for each intrinsic property, and there are as many types of stubs: each type of stubs corresponds to the type of the node that is at the other end of the edge. To simplify the notation, types of stubs will be identified by the type of the node toward which they point (i.e., $\mathcal{E} = \mathcal{N}$). Hence the joint degree distribution $P_i(\boldsymbol{k})$ prescribes the number of nodes of each type that nodes of type i are connected to. The conditions (4.2) ask that there are as many stubs stemming from nodes of type i toward nodes of type j as in the reverse direction, $w_i\langle k_j \rangle_{P_i} = w_j \langle k_i \rangle_{P_j}$. These constraints also prescribe the distribution

$$R(\boldsymbol{n}) = \frac{1}{R'} \sum_{i,j \in \mathcal{N}} (1 - \delta_{0,n_{ji}}) w_i \langle k_j \rangle_{P_i} \ , \tag{4.22}$$

where $R' = \sum_{i',j' \in \mathcal{N}} w_{i'} \langle k_{j'} \rangle_{P_{j'}}$ is simply the normalization factor. Assuming that nodes of type i exist with probability r_i, and that edges going from a node of type i to a node of type j exist with probability T_{ij} (i.e., edges may be more likely to exist in one direction that in the other), we get from Eqs. (4.1) and (4.4)

$$\theta_{ji}(\boldsymbol{x}) = (1 - r_j T_{ij}) + r_j T_{ij} x_{ij} \tag{4.23a}$$

$$\bar{\theta}_{ji}(\boldsymbol{x}) = (1 - r_j T_{ji}) + r_j T_{ji} x_{ij} \tag{4.23b}$$

for $i, j \in \mathcal{N}$. Using Eqs. (4.23) in Eqs. (4.5)–(4.15) and setting every $r_i = 1$ yields the results obtained in Ref. [10] for *multitype* graphs, which are themselves a generalization of several other formalisms [136, 137, 147].

An important category of correlations is the one based on the degree of nodes [135, 193]. These correlations are encoded in the conditional probability $P(d'|d)$ corresponding to the probability that the neighbor of a node with a degree d has a degree equal to d'. This can be reproduced with our formalism by considering every node with the same degree to be of the same type (i.e., a node of type i has i neighbors), and by using Eqs. (4.23) and the following degree distribution

$$P_i(\boldsymbol{k}) = \frac{i!}{\prod_{j'\in\mathcal{N}} k_{j'}!} \prod_{j\in\mathcal{N}} [P(j|i)]^{k_j} \,. \tag{4.24}$$

From Eq. (4.5), we obtain

$$g_i(\boldsymbol{x}) = \left[\sum_{j\in\mathcal{N}} P(j|i)\theta_{ji}(\boldsymbol{x})\right]^i \,, \tag{4.25}$$

and Eq. (4.6) yields

$$f_{vi} = \left[\sum_{j\in\mathcal{E}} P(j|i)\theta_{ji}(\boldsymbol{x})\right]^{i-1} \,, \tag{4.26}$$

which is independent of the type of the node/stub, namely v, from which the node has been reached. This is a direct consequence of the multinomial distribution in Eq. (4.24) and shows that our approach, through the distribution $P_i(\boldsymbol{k})$, can include more detailed correlations in the degree of the neighbors of nodes. Setting every $r_i = r$ and every $T_{\alpha i} = T$ in Eqs. (4.25)–(4.26), Eqs. (4.7)–(4.12) yield the results obtained in Refs. [135, 193]. Replacing $P(i|j)$ by $\frac{iw_i}{\sum_{i\in\mathcal{N}} jw_j}$ in Eq. (4.24) and letting the probability for nodes to exist to depend on the degree of the nodes then yields the results obtained in Ref. [37] which allows to study targeted removal of nodes based on their degree in uncorrelated random graphs. Finally, we have used this general approach in Ref. [92] to define an ensemble of graphs that reproduce the correlations induced by the k-core structure found in graphs extracted from real complex systems.

4.6.4 Clustered random graphs

We now turn our attention to random graphs models involving clustered hyperedges (i.e., hyperedges that contain loops). Most, if not all, variants of the CM containing clustered hyperedges are special cases of the approach presented in this paper. We illustrate this claim with a few examples.

Since the clustering property is related to the number of triangles found in graphs—hence capturing the idea that *the friend of my friend is also my friend*—it is natural to introduce clustering in graphs through the use of triangles (i.e., three nodes all connected together) [98, 127, 143, 177]. The simplest corresponding graph ensemble then has $|\mathcal{N}| = 1$ types of nodes (type 0, $w_0 = 1$), and $|\mathcal{E}| = 2$ types of stubs: two stubs of type A are paired to form undirected edges and three stubs of type B are matched to create triangles. Note that only one stub of type B is required to belong to a triangle even though its contribution amounts to two to the degree of the node; stubs can be seen as a membership to an hyperedge. The constraints given by Eq. (4.2) lead to $R(2,0) = 1 - R(0,3) = 3\langle k_A \rangle_{P_0} / (3\langle k_A \rangle_{P_0} + 2\langle k_B \rangle_{P_0})$. Assuming that nodes and edges exist with probabilities r and T, we obtain from Eqs. (4.1) and (4.4)

$$\theta_{A0} = (1 - rT) + rTx_{A0} \tag{4.27a}$$

$$\begin{aligned} \theta_{B0} &= (1 - rT)^2 + 2rT[1 - rT(2 - T)]x_{B0} \\ &\quad + r^2T^2[3 - 2T]x_{B0}^2 \, . \end{aligned} \tag{4.27b}$$

Using these two functions in Eqs. (4.5)—(4.15) leads directly to the results obtained in Ref. [98, 127, 143, 180]. Similarly, the results of Ref. [77] can be obtained with three types of stubs, $\mathcal{N} = \{1, 2, 3\}$, and one type of stubs, $\mathcal{E} = \{A\}$, where all hyperedges are triangles containing one node of each type [$R(1,1,1) = 1$ and $\theta_{Ai}(\boldsymbol{x}) = x_{A0}x_{A1}x_{A2}/x_{Ai}$].

Besides triangles, clustering—or any digression from a perfect tree-like structure—has been introduced in random graphs through the inclusion of various categories of hyperedges that involve more than three nodes. For instance, in Ref. [80, 84, 138] clustering is incorporated through fully connected hyperedges, or *cliques*, where nodes or edges exist with given probabilities (i.e., Erdős-Rényi graphs). In all cases, there is only one type of nodes. We retrieve the model of Ref. [138] by using one type of stubs; $P_0(k)$ prescribes the number of cliques to which nodes belong, and $R(n)$ prescribes the number of nodes in each clique. In the model considered in Refs. [80, 84], nodes belong to only one clique, but can have many single edges. As the number of single edges and the size of the clique can be correlated, there is one type of stubs for each clique size and an additional type for single edges; cliques of size m are formed by matching m stubs of the corresponding type. Hence the structure of the graphs is fully prescribed by $P_0(\boldsymbol{k})$ whose argument indicates the number of single edges and the size of the clique, and the constraints (4.2) yields

$$R(\boldsymbol{n}) = \frac{1}{R''} \sum_{\beta \in \mathcal{E}} (1 - \delta_{0,n_{\beta 0}}) \frac{\langle k_\beta \rangle}{n_{\beta 0}} \, , \tag{4.28}$$

where $R'' = \sum_{\beta \in \mathcal{E}} \langle k_\beta \rangle / n_{\beta 0}$ is the normalization constant. Using these distributions and quantities, our model reproduces the ones presented in Refs. [80, 84]. Also, we have used a version of our model that is similar to the one introduced in Ref. [84] to uncover a transition in the effectiveness of immunization strategies [91].

Finally, two of the most versatile models published to date are also special cases of our model. The one published in Ref. [6] is a previous version of the model presented in this paper. The two main differences are that the previous version did not handled site percolation, and that only stubs of same type could be matched to create hyperedges (e.g., no directed edge unless it is between nodes of different types). The model published in Ref. [98] can be retrieved from our model with one node type ($|\mathcal{N}| = 1$) and with one type of stubs

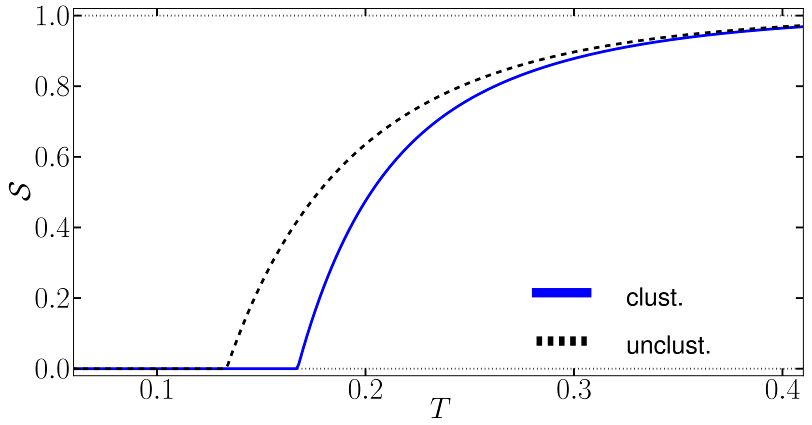


FIGURE 4.3 – Comparison of the emergence of the giant component in a clustered graph ensemble that qualifies for the strong clustering regime with its unclustered counterpart. We see that the latter has a larger giant component and a lower percolation threshold. Details on the graphs used are given in the main text.

for each *role* a node can play in hyperedges. This model can handle site and bond percolation but requires to solve percolation on each hyperedge beforehand, as in our model. While such calculation consists in a mere enumeration of each possible configuration of existing nodes and edges, it rapidly becomes cumbersome as the number of nodes and edges increases. Equations (4.1) then offers a systematic method to perform these calculations and therefore further increases the number of different configurations of hyperedges that can be handled analytically.

4.6.5 Weak and strong clustering regimes

We now use our model to test a conjecture regarding the effect of clustering (e.g., triangles) on bond percolation. References [177, 176] proposed that clustering has opposite effects on the bond percolation threshold and on the size of giant component depending of the density of triangles in a graph. This density is measured through the degree dependent clustering coefficient $\bar{c}(k)$: the probability that two neighbors of a node of degree k are also neighbors (i.e., they complete the triangle). The conjecture states that the *weak* clustering regime $\bar{c}(k) < (1-k)^{-1}$ leads to a higher percolation threshold and to a smaller giant component than in an equivalent unclustered graph. Contrariwise, *strong* clustering is defined such that $\bar{c}(k) > (1-k)^{-1}$ and leads to a lower percolation threshold and to a larger giant component than in an equivalent unclustered graph.

Let us consider the following graph ensemble in which there are two types of nodes $\mathcal{N} = \{0,1\}$ and three types of edges $\mathcal{E} = \{A, B, C\}$. Every node of type 0 has one stub of type A and one stub of type B, while each node of type 1 has one stub of type B and one stub of type C. Hyperedges are formed by matching either 4 stubs of type A, 4 stubs of type B (two stemming from each type of nodes), or 8 stubs of type C; nodes are all connected to one another in every hyperedges. The constraints given by Eq. (4.2) imply that $w_0 = w_1$ and that $R(\boldsymbol{n}) = R(n_{A0}, n_{B0}, n_{B1}, n_{C1})$ follows the relation $2R(4,0,0,0) = R(0,2,2,0) = 4R(0,0,0,8) = 4/7$. Nodes of type 0 therefore have a degree of 6 and $\bar{c}(6) = \frac{6}{15} > \frac{1}{5}$, and nodes of type 1 have a degree equal to 10 and $\bar{c}(10) = \frac{24}{45} > \frac{1}{9}$. This graph ensemble clearly qualifies for the strong regime.

To isolate the effect of clustering on bond percolation, we compare the results obtained for the graph ensemble described above with the ones obtained with an equivalent unclustered version. This equivalent graph ensemble possesses identical correlations, but hyperedges are broken into individual independent edges instead. The behavior of its giant component is obtained as in Sec. 4.6.3 with $P_0(4,2) = P_1(2,8) = 1$.

Figure 4.3 compares the behavior of the giant component in both ensembles when edges exist with probability T. We conclude that although the clustered graph ensemble qualifies for the strong regime, the behavior observed is the one of the weak regime: higher percolation threshold and smaller giant component than for the equivalent unclustered graph. This behavior can be understood in terms of branching factors. The unclustered graphs have a tree-like structure and therefore maximize the number of *new* nodes encountered while navigating the graph: every edge leads to a new node. The redundancy caused by clustering means that not all edges lead to a new node in the clustered graphs, which reduces the average number of nodes that can be reached from any given node. Hence a larger number of edges must exist for a giant component to appear (e.g., larger threshold), and this component will be smaller as many edges will be *wasted* by leading to nodes previously reached.

This counterexample strongly suggests that the criterion on $\bar{c}(k)$ could be a necessary condition for a strong clustering regime but that it is not a sufficient one. The explanation in terms of branching factors alongside the results in Refs. [81, 103, 127] point toward the conclusion that the behavior of graphs with an *underlying* tree-like structure is best described by the weak clustering regime.

4.6.6 Bijection between site and bond percolation thresholds

From Eqs. (4.6), we see that the elements of the Jacobian matrix $\mathbf{J}$ used to determine the point at which the giant component appears have the general form

$$\frac{\partial f_{\mu i}(\mathbf{1})}{\partial x_{\nu j}} = \sum_{\alpha \in \mathcal{E}} \frac{\langle k_\mu (k_\alpha - \delta_{\mu\alpha}) \rangle_{P_i}}{\langle k_\mu \rangle_{P_i}} \frac{\partial \theta_{i\alpha}(\mathbf{1})}{\partial x_{\nu j}} \,, \tag{4.29}$$

for every $i, j \in \mathcal{N}$ and $\mu, \nu \in \mathcal{E}$. These terms are in fact branching factors: each element is the average number of nodes of type j reached by their stub of type ν that are present in the neighborhood of a node of type i that has itself being reached by its stub of type μ. More precisely, the first term corresponds to the average number of stubs of type α that a node of type i has if it has been reached from one of its stubs of type μ (this stub is excluded from the count if $\alpha = \mu$). The second term is the average number of pairs (ν, j) that are reached in hyperedges entered via a stub of type α of a node of type i. The value of these latter terms depends on the structure of hyperedges and on the probabilities for nodes and edges to exist in them.

Let us assume that all hyperedges have the same structure; nodes of different types may be involved in a nontrivial manner as long as all hyperedges have the same shape (e.g., they all are triangles). We also suppose that nodes and edges exist with probabilities that are independent of their type, namely r and T. In such case, every nonzero $\frac{\partial \theta_{i\alpha}(\mathbf{1})}{\partial x_{\nu j}}$ is a polynomial in r and T, $h(r, T)$, and is independent of i, j, α and ν. Consequently the dependency in r and T can be factored out of the Jacobian matrix

$$\mathbf{J} = h(r, T)\mathbf{J}' \ . \tag{4.30}$$

Since the giant component appears when $\lambda_{\max}(\mathbf{J}) = h(r, T)\lambda_{\max}(\mathbf{J}') = 1$, the points (r', T') at which the phase transition occurs all belong to the critical surface

$$h(r', T') = \frac{1}{\lambda_{\max}(\mathbf{J}')} \ . \tag{4.31}$$

Whenever applicable, this result relates a point (r_1, T_1) at which the graphs are known to percolate to any other critical point (r_2, T_2) through $h(r_1, T_1) = h(r_2, T_2)$. For instance, this relation leads to a direct bijection between the thresholds of pure site percolation $(r_c, 1)$ and pure bond percolation $(1, T_c)$ through $h(r_c, 1) = h(1, T_c)$.

4.6.7 Emergence of the functional component

As a final example, we briefly investigate the effect of clustering and dependency edges on the size of the extensive component. To do so, we first consider the edge-triangle clustered graph ensemble (see Sec. 4.6.4) with the joint degree distribution $P_0(0, 3) = P_0(2, 1) = 2P_0(2, 0) = 4/10$. Notice that this joint degree distribution forces assortative mixing as high and low degree nodes tend to be segregated. The size of the extensive component is given by the red curve in Fig. 4.4 as a function of the node existence probability r.

To illustrate the impact of dependency edges, we consider the case of two identical fully-dependent graphs (every node has a twin node, $q_{00}^{AB} = q_{00}^{BA} = 1$) where both graphs are identical to the one described above. The size of the extensive functional component in both graphs ($\mathcal{S}^A = \mathcal{S}^B = \mathcal{S}$) is shown in blue in Fig. 4.4. As expected, dependency edges cause the extensive functional component to appear through a discontinuous transition, whereas the extensive component in the isolated graphs emerges continuously. Figure 4.4 also shows that the extensive functional component is smaller than the giant component in the corresponding isolated graphs. This is in fact always the case since being part of a functional

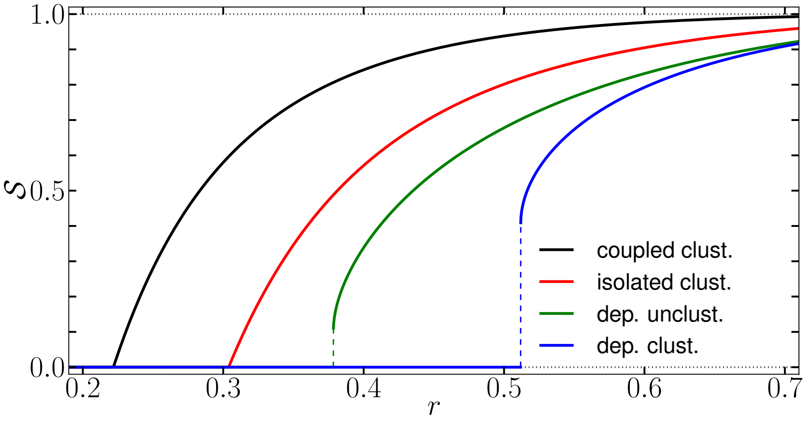


FIGURE 4.4 – Comparison of the size of the extensive component as a function of the node existence probability r on four related graph ensembles. The details of each graph ensembles are given in the main text.

component requires twin nodes to belong to the extensive component in their respective graphs. The size of the extensive functional component is therefore bounded by the size of the giant component.

To investigate the effect of clustering on the emergence of the extensive functional component, we consider the unclustered version of the two identical fully-dependent graphs in which triangles are broken down into two independent single edges. To preserve the correlations present in the clustered dependent graphs, two types of stubs are still used to distinguish original single edges from single edges that used to be triangles [127]. The extensive functional component in this graph ensemble appears at a lower value of r (in green, Fig. 4.4) than in the clustered ensemble. Since clustering raises the percolation threshold in isolated random graphs, this behavior was expected for the reason outlined at the end of the last paragraph. What was perhaps less expected is the effect of clustering on the jump size at the transition that our results suggest. Whether clustering systematically increases the jump size at the transition is the subject of an ongoing investigation and will be addressed in a future publication.

Finally, to further highlight the impact of dependency edges, we consider the case where two clustered graphs (defined above) are coupled with directed edges to mimic the dependency relationships found in interdependent graphs. Two types of nodes are used to distinguish the nodes from the two graphs, and a directed edge runs from the *influential* node to the *dependent* node (whose state depends on the state of the influential node). Hence, each node has one incoming directed edge from a node of the other type to account for the influence of the twin node. Additionally, each node also has a Poisson distributed (mean value 1) number of outgoing directed edges toward nodes of the other type since dependency edges are attributed randomly (as in the two identical fully-dependent graphs considered above). The size of the extensive component is shown in black in Fig. 4.4. The transition is of course continuous and occurs at a lower value than for the isolated clustered graphs. This example emphasizes the difference of role between dependency and connectivity edges. The former restrict the emergence of an extensive component as they requires twins nodes to belong to the extensive component in their respective graph, while the latter enhance the emergence of an extensive component as they offer new pathways for an extensive component to emerge. It is therefore not surprising that the coupled clustered graphs percolate at lower values of r than the two identical fully-dependent graphs.

4.7 Conclusion

Based on some of our previous work [6, 7, 10, 91, 92], we have presented a unifying conceptual framework that offers a comprehensive mathematical description of a wide variety of structural properties found in graphs extracted from real complex systems (e.g., correlations, segregation, clustering of various forms). The generality of this framework resides on the use of a multitype perspective that permits a precise prescription on how nodes are connected to one another, and on the use of a set of iterative equations that solve the distribution of the size of components in small arbitrary graphs. These iterative equations for solving the distribution of the size of components in finite-size arbitrary graphs are by themselves a valuable addition to graph theoretical methodology. Besides being a cornerstone of our formalism by allowing a mapping of hyperedges unto an effective tree-like structure, they have potential applications in the theoretical description of fragmentation processes [53, 54], and in "traditional" Percolation Theory [165, 172].

Our approach leads to the definition of a very general random graph ensemble for which site and/or bond percolation can be solved exactly using probability generating functions in the thermodynamical limit (e.g., size of the giant component, percolation threshold, distribution of the size of small components). We have shown that this random graph ensemble encompasses most, if not all, random graph models published to this day and that it can in fact include structural properties that were yet to be included in a theoretical framework. This versatility makes it a perfect theoretical laboratory that can be used to investigate the role of specific local structural properties on the global connectivity of the graphs. We have provided an example of this claim by using our mathematical approach to give a counterexample to a conjecture on the effect of clustering on the size of the giant component and on the percolation threshold. This counterexample in turn points toward a rewording of the conjecture and offers clues for future investigations.

Our formalism is also naturally well-suited for the modeling of interdependent graphs whose most striking feature is the emergence of the extensive component via a discontinuous transition. We have shown how the mathematical tools developed for independent graphs can be applied to interdependent graphs: allowing, for the first time, the inclusion of clustering in an interdependent graph model. We have presented a simple example that suggested that not only does clustering increase the percolation threshold, as expected, but that it also increases the size of the discontinuity in the emergence of the extensive component.

Percolation is a simple theoretical concept that focuses on the effect of local structural properties on the global connectivity of graphs. As such, it helps reveal the interplay between the underlying structure of interactions in complex systems and their functions. By offering the most comprehensive mathematical description of continuous and discontinuous percolation on random graphs to date, we believe that this work will contribute to a better understanding of complex systems.

Chapitre 5

Application I : Coexistence de composantes extensives dans des graphes aléatoires

Article original :

Coexistence of phases and the observability of random graphs

Antoine Allard[1], Laurent Hébert-Dufresne, Jean-Gabriel Young et Louis J. Dubé

Département de Physique, de Génie Physique, et d'Optique,
Université Laval, Québec (Qc), Canada G1V 0A6

Référence : Référence : Phys. Rev. E, **89** 022801 (2014)[2] [9]

© 2013 American Physical Society (§5.2 –§5.8)[3]

[1] Le développement théorique, les simulations numériques de même que la rédaction du manuscrit sont essentiellement les fruits du travail du premier auteur. Les autres auteurs ont participé à l'élaboration générale du projet de même qu'à la révision du manuscrit.

[2] Paru dans la section *Editors' Suggestions* de Phys. Rev. E.

[3] Ces sections contiennent le contenu original de l'article. Celui-ci n'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

Cet article est un exemple probant de la nature imprévisible de la recherche scientifique. Le tout débuta par la lecture d'un article de l'équipe de Adilson E. Motter (Northwestern University) sur l'*observabilité* des systèmes complexes [200]. L'objectif était alors simplement d'en faire la présentation aux membres du groupe lors d'une de nos réunions hebdomadaires ; la présentation d'articles étant une activité régulière dans le groupe de recherche de Louis J. Dubé. L'idée explorée dans l'article était originale, mais l'approche mathématique pour en investiguer les tenants et aboutissants était inutilement compliquée et peu intuitive. Ceci était d'autant plus vrai lorsque celle-ci était considérée à travers la lunette de l'approche multitype. Il fut donc entrepris de traduire leur formalisme en ces termes et, à cet égard, la version multitype de l'ensemble CM était l'outil adéquat [5, 10].

Non seulement fut-il possible de démontrer que leur formalisme pouvait être traduit de façon exacte en termes de l'approche multitype, mais également que le nouveau formalisme permettait d'approfondir de façon significative l'investigation initiée par l'équipe du Prof. Motter. D'une part, notre approche est beaucoup plus intuitive : les objets mathématiques possédent désormais une interprétation « physique » claire et directe. Notre approche est également systématique. Ceci nous permit de développer un formalisme général (en l'occurrence pour toutes profondeurs L), alors que l'approche précédente requérait un nouvel ensemble d'équations pour chaque valeur de L. Notre approche nous permit même de mettre au jour une « erreur » (du moins une approximation aucunement mentionnée) dans leur formalisme pour $L = 2$. D'autre part, notre approche mena à la démonstration d'une possible coexistence de deux composantes extensives dans un même graphe — fait rarissime en théorie de la percolation — et permit de tirer les conclusions appropriées à propos de l'observabilité de systèmes réels.

Enfin, en comparant les prédictions de notre formalisme avec les résultats obtenus à l'aide de graphes extraits de systèmes complexes réels, nous fûmes en mesure de caractériser la structure de ces graphes. Nos résultats suggèrent en effet qu'en ce qui a trait à leur observabilité, les systèmes réels peuvent être distingués en deux catégories selon qu'ils soient contraints dans un espace physique (c.-à-d. un espace euclidien de deux ou trois dimensions) ou non. Quoique préliminaire, cette classification propose des avenues intéressantes pour améliorer et rendre encore plus réaliste la description mathématique de systèmes complexes.

Quantity	Description	Definition
L	Depth of the percolation/observability	Sec. 5.4
φ	Probability for a node to be directly occupied/monitored	Sec. 5.4
M	Number of types of nodes	Sec. 5.4.2
$P(k)$	Degree distribution	Sec. 5.4
$G_0(x)$	pgf generating the degree distribution	Sec. 5.4
$G_1(x)$	pgf generating the excess degree distribution	Sec. 5.4
$\langle k \rangle$	Average degree of nodes	Sec. 5.4
$P_i(\boldsymbol{k})$	Joint degree distribution of nodes of type i	Sec. 5.4.1
$g_i^{(L)}(\boldsymbol{x})$	pgf generating the joint degree distribution $P_i(\boldsymbol{k})$	Eq. (5.21)
$w_i^{(L)}$	Fraction of nodes that are of type i	Eq. (5.20)
$\varepsilon_i^{(L)}$	Probability that a randomly chosen edge leads to a type i node	Sec. (5.18)
$\chi_i^{(L)}$	Probability that the type of the node at the end of a random edge is not lower than i	Eq. (5.19)
$a_{ij}^{(L)}$	Probability for an edge that leaves a node of type i towards a node of type j to not lead to an extensive component	Eqs. (5.23) and (5.25)
$S^{(L)}$	Size of the giant occupied/observable component	Eq. (5.22)
$\varphi_c^{(L)}$	Value of φ at which point the giant occupied/observable component emerges	Sec. 5.5
$\bar{S}^{(L)}$	Size of the giant non-occupied/non-observable component	Eq. (5.24)
$\bar{\varphi}_c^{(L)}$	Value of φ at which point the giant non-occupied/non-observable component emerges	Eq. (5.26)

TABLE 5.1 – Glossary of the major mathematical objects defined in this chapter.

5.2 Abstract

In a recent Letter, Yang *et al.* [Phys. Rev. Lett. **109**, 258701 (2012)] introduced the concept of *observability transitions*: the percolation-like emergence of a macroscopic *observable* component in graphs in which the state of a fraction of the nodes, and of their first neighbors, is monitored. We show how their concept of depth-L percolation—where the state of nodes up to a distance L of monitored nodes is known—can be mapped unto multitype random graphs, and use this mapping to exactly solve the observability problem for arbitrary L. We then demonstrate a non-trivial coexistence of an observable and of a non-observable extensive component. This coexistence suggests that monitoring a macroscopic portion of a graph does not prevent a macroscopic event to occur unbeknown to the observer. We also show that real complex systems behave quite differently with regard to observability depending on whether they are geographically-constrained or not.

5.3 Introduction

Considered as the ultimate proof of our understanding, the *controllability* (and its dual concept the *observability*) of natural and technological complex systems have been the subject of many recent studies [47, 96, 97, 111, 112, 113, 134, 160, 167, 168, 189, 199, 200]. In essence, the question is whether the global state of a system can be imposed (inferred) through the control (monitoring) of a few of its constituents. By mapping the underlying web of interactions between the constituents of systems unto graphs, analytical criteria have been proposed to determine whether a system is controllable (observable), and if so, through which of its constituents control (monitoring) should be applied. However, although promising and theoretically correct, doubts have been raised as to whether these criteria can be used in practice on real large systems [47, 189].

Whenever a comprehensive and exact theoretical framework is lacking, simpler but solvable theoretical models—that consider simplified versions of the systems under scrutiny—become valuable alternatives to highlight and understand key behaviors of complex systems. Following this trend, Yang *et al.* used random graphs to study the observability of power grids through the use of *phasor measurement units* that allow to monitor the state of a node and of each of its neighbors [200]. Using this approach, they demonstrated that the largest observable component emerges in a percolation-like transition, and argued that structural properties found in real systems reduce the number of monitoring units required for achieving large-scale observability.

We formalize their approach into the general concept of depth-L percolation where the state of nodes up to a distance L of monitored nodes is known as well. Using a multitype version of the Configuration Model [10], we study analytically the emergence of the extensive "giant" observable component (i.e., its size and the conditions for its existence), and we

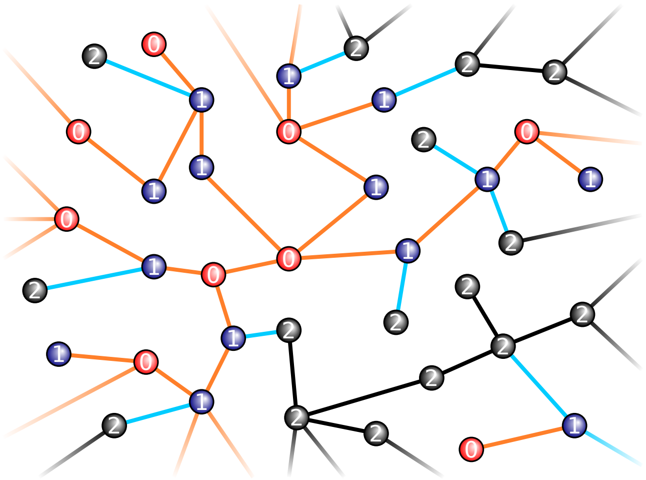


FIGURE 5.1 – Illustration of depth-1 percolation on a graph generated through the Configuration Model. Directly occupied, indirectly occupied and non-occupied nodes are in red (type 0), blue (type 1) and black (type 2), respectively. Occupied components are identified with orange edges, and non-occupied components with black edges. Edges linking occupied and non-occupied components—the ones removed by setting $x_{02} = x_{12} = x_{20} = x_{21} = 1$ in Eqs. (5.4), (5.8) and (5.9)—are shown in cyan.

demonstrate a non-trivial coexistence with another extensive component: one made of non-monitored nodes. We then turn our attention to graphs extracted from real complex systems and show that many such systems support the coexistence of two extensive components. Moreover, our theoretical framework yields analytical arguments to explain the low thresholds for the large-scale observability observed in many of these systems. However, we find that geographically-constrained systems (e.g., power grids) are poorly modeled by random graphs; rather, their topology appears similar to the one of lattices. Our results also suggest that they behave quite differently with regard to observability: their structure does not support coexistence, and achieving large-scale observability requires more monitoring units than hinted by calculations based on the Configuration Model [200].

This paper is organized as follows. In Sec. 5.4, we introduce the concept of depth-L percolation and develop an exact mathematical description for the case $L = 1$. This allows one to demonstrate the equivalence between our approach and the one proposed in Ref. [200], and to identify the possible coexistence of two extensive components. In Sec. 5.5, we generalize our mathematical framework to any L, and use it to study the effect of varying the depth on the coexistence regime. In Sec. 5.6, we investigate the observability of graphs extracted from real complex systems with numerical simulations and our mathematical framework. Conclusions and final remarks are collected in the last section. A technical Appendix is supplied to describe the case $L = 2$ and to compare it with the results obtained in Ref. [200].

5.4 Depth-L percolation

Depth-L percolation is a generalization of traditional site percolation: nodes are occupied independently with probability φ *and* every node up to a distance L of occupied nodes are also occupied. We say that the latter are *indirectly occupied* as opposed to the former which are said to be *directly occupied* (see Fig. 5.1). Depth-0 percolation corresponds to traditional site percolation (see Sec. 5.5.1). For the sake of simplicity (and to make an explicit correspondence with the mathematical treatment in Ref. [200]), we first focus on depth-1 percolation—where first neighbors of occupied nodes are occupied as well—on graphs generated through the Configuration Model [144]. The generalization to any L is however straightforward in our formalism and is the subject of Sec. 5.5.

The Configuration Model defines a maximally random graph ensemble whose graphs are random in all respects except for the degree distribution, $\{P(k)\}_{k\in\mathbb{N}}$, prescribing the number of connections that nodes have (i.e., number of first neighbors). Using probability generating functions (pgf), many exact results can be obtained in the limit of large graphs [6, 136, 147]. For the present study, we define the pgf associated with the degree distribution

$$G_0(x) = \sum_{k=0}^{\infty} P(k)x^k \;, \tag{5.1}$$

and the one generating the number of edges *leaving* a node reached by one of its edges (*excess* degree distribution)

$$G_1(x) = \frac{G_0'(x)}{G_0'(1)} = \frac{1}{\langle k\rangle}\sum_{k=1}^{\infty} kP(k)x^{k-1} \;. \tag{5.2}$$

Here the prime denotes the derivative, and $\langle k\rangle$ corresponds to the first moment of the degree distribution (i.e., average degree). The Configuration Model generates graphs through a stub pairing scheme: a random number of stubs (the degree) is assigned to each node according to $\{P(k)\}_{k\in\mathbb{N}}$, and edges are formed by randomly matching stubs together. In the context of depth-L percolation, directly occupied nodes are then selected with probability φ, and the identification of indirectly occupied nodes follows.

5.4.1 Mapping to multitype random graphs

To study the emergence of the extensive occupied component, we introduce a mapping linking depth-L percolation to percolation on multitype random graphs [10]. To facilitate this mapping, we consider an alternative procedure to generate graphs with directly and indirectly occupied nodes. As previously stated, a degree is assigned to each node according to $\{P(k)\}_{k\in\mathbb{N}}$, but instead of pairing stubs right away, directly occupied nodes (type 0) are first selected with probability φ. Stubs of type 0 nodes are then randomly matched with any stubs in the graph; untagged nodes now connected to type 0 nodes are said to be indirectly occupied (type 1). Note that two type 0 nodes can be linked together. Nodes that have neither been tagged as type 0 nor as type 1 are said to be non-occupied (type 2). All remaining free stubs are finally paired randomly to *close* the graph. This alternative perspective is identical to the one discussed in the previous section in the limit of large graphs, and is analog to *on-the-fly* network construction [149]. Although it may seem unnecessary in the simple case $L = 1$, this slight change of perspective greatly eases the generalization to an arbitrary value of L.

By definition, a fraction $w_0 = \varphi$ of the nodes is of type 0. Because these nodes are assigned randomly and independently, the distribution of the number of connections they have with other node types (their *joint* degree distribution) is

$$P_0(\boldsymbol{k}) = \delta_{0k_2} P(k_0 + k_1) \frac{(k_0 + k_1)!}{k_0! k_1!} \varphi^{k_0} (1 - \varphi)^{k_1} \ , \tag{5.3}$$

where $\boldsymbol{k} = (k_0, k_1, k_2)$ and δ_{ab} is the Kronecker delta. In other words, if a neighbor of a type 0 node is not of type 0, it is inevitably of type 1. The associated pgf is

$$\begin{aligned} g_0(\boldsymbol{x}) &= \sum_{\boldsymbol{k}} P_0(\boldsymbol{k}) x_{00}^{k_0} x_{01}^{k_1} x_{02}^{k_2} \\ &= G_0\big(\varphi x_{00} + [1 - \varphi] x_{01}\big) \ . \end{aligned} \tag{5.4}$$

A randomly chosen node will be of type 1 if it has not been selected as a type 0 node and if at least one of its neighbors is of type 0. This happens with probability $(1 - \varphi)[1 - (1 - \varphi)^k]$ for a node whose degree is equal to k. Averaging over the degree distribution, we find

$$w_1 = (1 - \varphi)[1 - G_0(1 - \varphi)] \ . \tag{5.5a}$$

Asking for normalization, we find that type 2 nodes represent a fraction

$$w_2 = (1 - \varphi) G_0(1 - \varphi) \tag{5.5b}$$

of the nodes. Likewise, we define ε_i as the probability that a randomly chosen edge leads to a type i node. Clearly $\varepsilon_0 = \varphi$, and by similar arguments as above but by averaging over the excess degree distribution instead, we find that

$$\varepsilon_1 = (1 - \varphi)[1 - G_1(1 - \varphi)] \tag{5.6a}$$

$$\varepsilon_2 = (1 - \varphi) G_1(1 - \varphi) \ . \tag{5.6b}$$

From the alternative procedure described above, we find that the joint degree distribution of the union of type 1 and type 2 nodes is

$$P_{1 \bigcup 2}(\boldsymbol{k}) = P(k_0 + k_1 + k_2) \frac{(k_0 + k_1 + k_2)!}{k_0! k_1! k_2!} \varepsilon_0^{k_0} \varepsilon_1^{k_1} \varepsilon_2^{k_2} \ . \tag{5.7}$$

Since the difference between nodes of these two types is the presence of type 0 nodes in their immediate neighborhood, we can readily write the pgf associated with their joint degree distribution

$$
\begin{aligned}
g_1(\boldsymbol{x}) &= A_1 \sum_{\boldsymbol{k}} (1-\delta_{0k_0}) P_{1\bigcup 2}(\boldsymbol{k}) x_{10}^{k_0} x_{11}^{k_1} x_{12}^{k_2} \\
&= \frac{G_0\big(\varepsilon_0 x_{10} + \varepsilon_1 x_{11} + \varepsilon_2 x_{12}\big) - G_0\big(\varepsilon_1 x_{11} + \varepsilon_2 x_{12}\big)}{1 - G_0\big(1-\varphi\big)}
\end{aligned}
\tag{5.8}
$$

and

$$
\begin{aligned}
g_2(\boldsymbol{x}) &= A_2 \sum_{\boldsymbol{k}} \delta_{0k_0} P_{1\bigcup 2}(\boldsymbol{k}) x_{20}^{k_0} x_{21}^{k_1} x_{22}^{k_2} \\
&= \frac{G_0\big(\varepsilon_1 x_{21} + \varepsilon_2 x_{22}\big)}{G_0\big(1-\varphi\big)} \ ,
\end{aligned}
\tag{5.9}
$$

where A_1 and A_2 are normalization constants. With $\{w_i\}_{i=0,1,2}$ and the pgf $\{g_i(\boldsymbol{x})\}_{i=0,1,2}$ in hand, we are now in a position to mathematically describe the emergence of the giant occupied component.

5.4.2 Giant occupied component

It has been shown in Ref. [10] that the relative size of the giant component, $\mathcal{S}$, in multitype random graphs is computed via

$$
\mathcal{S} = \sum_{i=0}^{M-1} w_i \big[1 - g_i(\boldsymbol{a})\big]
\tag{5.10}
$$

where M is the number of node types, and where $\boldsymbol{a} = \{a_{ij}\}_{i,j=0,\ldots,M-1}$ is the set of probabilities that an edge leaving a type i node towards a type j node *does not* lead to the giant component. These probabilities correspond to the stable fixed point—the smallest solution in $[0,1]^{M^2}$—of the following system of equations

$$
a_{ij} = \frac{\partial g_j(\boldsymbol{a})/\partial x_{ji}}{\partial g_j(\mathbf{1})/\partial x_{ji}}
\tag{5.11}
$$

with $i, j = 0, \ldots, M-1$. We are interested in the relative size of the giant occupied component, the component made of type 0 and type 1 nodes solely. To do so, we isolate them from type 2 nodes by setting $x_{02} = x_{12} = x_{20} = x_{21} = 1$ in Eqs. (5.4), (5.8) and (5.9). Noting that $a_{10} = a_{00}$, this yields

$$
a_{00} = G_1\big(\varphi a_{00} + (1-\varphi) a_{01}\big)
\tag{5.12a}
$$

$$
a_{01} = G_1\big(\varphi a_{00} + \varepsilon_1 a_{11} + \varepsilon_2\big)
\tag{5.12b}
$$

$$
a_{11} = \frac{G_1\big(\varphi a_{00} + \varepsilon_1 a_{11} + \varepsilon_2\big) - G_1\big(\varepsilon_1 a_{11} + \varepsilon_2\big)}{1 - G_1\big(1-\varphi\big)}
\tag{5.12c}
$$

and the relative size of the giant occupied component, S, becomes [summing Eq. (5.10) only over $i = 0, 1$]

$$\begin{aligned} S = 1 - \varphi G_0\big(\varphi a_{00} + (1-\varphi)a_{01}\big) \\ - (1-\varphi)\Big\{G_0\big(\varphi a_{00} + \varepsilon_1 a_{11} + \varepsilon_2\big) \\ + G_0\big(1-\varphi\big) - G_0\big(\varepsilon_1 a_{11} + \varepsilon_2\big)\Big\} \,. \end{aligned} \tag{5.13}$$

Clearly, $a_{00} = a_{01} = a_{11} = 1$ is always a solution of Eqs. (5.12) and corresponds to the situation where there is no giant occupied component ($S = 0$). A giant occupied component emerges in fact when the stable fixed point $\boldsymbol{a} = \mathbf{1}$ undergoes a transcritical bifurcation during which a stable fixed point appears in $[0,1)^3$. Hence a linear stability analysis of $\boldsymbol{a} = \mathbf{1}$ leads to the criterion

$$\begin{aligned} G_1'(1) = 1 + (1-\varphi_c)G_1'(1-\varphi_c) \\ \times \left\{1 - \varphi_c G_1'(1) - \varphi_c(1-\varphi_c)\big[G_1'(1)\big]^2\right\} \end{aligned} \tag{5.14}$$

marking the point $\varphi = \varphi_c$ where the giant occupied component starts to emerge. This is the exact same criterion obtained by Yang *et al.* [200]. In fact, by identifying $u \equiv \varphi a_{00} + (1-\varphi)a_{01}$ and $(1-\varphi)s \equiv \varepsilon_1 a_{11} + \varepsilon_2$, Eqs. (5.12)–(5.13) fall back on their results, thereby demonstrating the equivalence between the two approaches. Notice also that we retrieve from Eqs. (5.12)–(5.14) the well-known results for the Configuration Model [147] in the limit $\varphi \to 1$ (see the caption of Fig. 5.2).

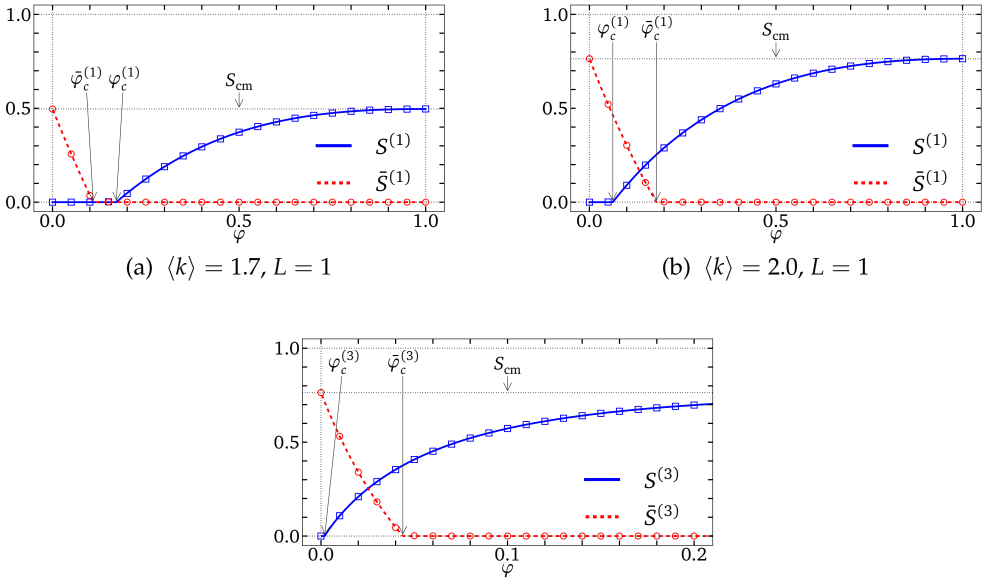
The multitype perspective offers a simple interpretation of the emergence of the giant occupied component. For such component to exist, the original graph ensemble—defined by $\{P(k)\}_{k\in\mathbb{N}}$—must itself have a giant component, which occurs when $G_1'(1) > 1$ [147]. A giant *occupied* component then exists if an extensive component composed of only type 0 and type 1 nodes prevails after edges between occupied and non-occupied nodes have been removed (cyan edges in Fig. 5.1). In other words, a giant occupied component exists if the original giant component is robust to the removal of these edges. Yet if the original giant component is *robust enough*, an extensive component composed of non-occupied nodes solely could also prevail, therefore leading to the coexistence of two extensive components.

5.4.3 Giant non-occupied component

One equation has been left out of Eqs. (5.12). Indeed, Eqs. (5.11) yields another nontrivial equation

$$a_{22} = \frac{G_1\big(\varepsilon_1 + \varepsilon_2 a_{22}\big)}{G_1\big(1-\varphi\big)} \tag{5.15}$$

for the probability that an edge between two type 2 nodes does not lead to an extensive component. Since this component is made of non-occupied nodes solely, we refer to it as the



(a) $\langle k \rangle = 1.7, L = 1$

(b) $\langle k \rangle = 2.0, L = 1$

(c) $\langle k \rangle = 2.0, L = 3$

FIGURE 5.2 – Validation of the theoretical formalism for different depth of percolation (L). Size of the occupied and non-occupied components in function of φ for graph ensembles with a different average degree. The degrees of both ensembles are exponentially distributed according to $P(k) = (1 - \mathrm{e}^{-\lambda})\mathrm{e}^{-\lambda(k-1)}$ with $k \geq 1$ and $\lambda = -\ln\big(1 - 1/\langle k \rangle\big)$. The size of the giant component in the original graph ensemble, S_{cm}, is shown for comparison. It is equal to $S_{\mathrm{cm}} = 1 - G_0(a)$ where a is the solution of $a = G_1(a)$ [147]. Curves are the solutions of Eqs. (5.12)–(5.13) and (5.15)–(5.16) [$L = 1$], and Eqs. (5.22)–(5.23) and (5.24)–(5.25) [$L = 3$]. Symbols are the relative size of the largest occupied and non-occupied components averaged over at least 100 graphs of at least 5×10^5 nodes each. Threshold values were obtained from Eqs. (5.14), (5.17) and (5.26), and by analyzing the stability of Eqs. (5.23) around $\boldsymbol{a} = \mathbf{1}$. Note the change of scale of the abscissa of (c).

giant non-occupied component. By summing Eq. (5.10) over type 2 nodes only, the relative size of this other extensive component is

$$\bar{S} = w_2 \left[1 - \frac{G_0\big(\varepsilon_1 + \varepsilon_2 a_{22}\big)}{G_0\big(1 - \varphi\big)} \right] . \tag{5.16}$$

Again, we see that $a_{22} = 1$ is always a solution of Eq. (5.15) and the point $\varphi = \bar{\varphi}_c$ at which it becomes an unstable fixed point, that is when

$$(1 - \bar{\varphi}_c)G_1'\big(1 - \bar{\varphi}_c\big) = 1 \, , \tag{5.17}$$

marks the (dis)appearance of the giant non-occupied component. Again, notice that Eqs. (5.15)–(5.17) fall back on the results for the Configuration Model [147] in the limit $\varphi \to 0$ (see the caption of Fig. 5.2).

5.4.4 Coexistence of extensive components

Figure 5.2 depicts the typical scenarios with respect to the coexistence of two extensive components. In Fig. 5.2a, the size of the giant non-occupied component—initially equal to the size of the original giant component S_{cm}—decreases with increasing φ until the component stops being extensive at $\varphi = \bar{\varphi}_c$. Then there is an interval $[\bar{\varphi}_c, \varphi_c]$ where there is no extensive component: the whole graph is fragmented into non-extensive observable islands. The giant occupied component finally emerges at $\varphi = \varphi_c$ and its size increases with increasing φ until it is equal to the size of the original giant component. The same behavior is observed in Fig. 5.2b except that in this case the original giant component is dense enough for the giant occupied component to emerge *before* the giant non-occupied component disappears. Hence whenever $\varphi_c < \bar{\varphi}_c$, there is an interval $[\varphi_c, \bar{\varphi}_c]$ in which two extensive components coexist.

In the context of observability as considered by Yang *et al.*, directly occupied nodes are monitored in such a way that the state of their first neighbors is known as well (case where $L = 1$, see the Appendix for a discussion of the case $L = 2$) [200]. The existence of a giant occupied component then means that a macroscopic contiguous fraction of the graph can be monitored. However, coexistence suggests that monitoring a macroscopic portion of a graph does not prevent a macroscopic event to occur on this graph unbeknown to the observer. The condition for which there is coexistence is rather simple: the underlying extensive component (the one of the original graph) must be *sufficiently dense* to sustain two giant components. As discussed in Sec. 5.6, this condition is fulfilled in several real systems, with coexistence extending over a wide interval $[\varphi_c, \bar{\varphi}_c]$ in some cases.

5.5 Mathematical description for arbitrary depth

The mapping to multitype random graphs can be readily generalized to an arbitrary depth (L). The procedure to generate these graphs proceeds initially as for $L = 1$, but instead of closing the graph after type 1 nodes have been selected, the remaining free stubs stemming out of type 1 nodes are randomly paired with any free stubs in the whole graph. The nodes thereby reached have either already been tagged as type 1, or have not been tagged and are henceforth considered to be of type 2. The remaining free stubs of type 2 nodes are then randomly paired with any free stubs in the whole graph to determine type 3 nodes. This iterative assignment of node types is repeated until type L nodes are selected. The graph is then finally closed by randomly matching all remaining free stubs; nodes that have not been assigned a type are said to be non-occupied (type $L + 1$). In the end, there is a total of $L + 2$ node types.

With this iterative assignment of node types in mind, we generalize the mathematical description introduced in the previous section. The probability $\varepsilon_i^{(L)}$ that a random edge leads to a type i node is

$$\varepsilon_i^{(L)} = \begin{cases} \varphi & i=0 \\ (1-\varphi)\big[G_1\big(\chi_{i-1}^{(L)}\big) - G_1\big(\chi_i^{(L)}\big)\big] & 1 \leq i \leq L \\ (1-\varphi)G_1\big(\chi_L^{(L)}\big) & i = L+1 \end{cases} , \tag{5.18}$$

where we have defined

$$\chi_i^{(L)} = \begin{cases} 1 & i=0 \\ 1-\sum_{j=0}^{i-1}\varepsilon_j^{(L)} & i \geq 1 \end{cases} . \tag{5.19}$$

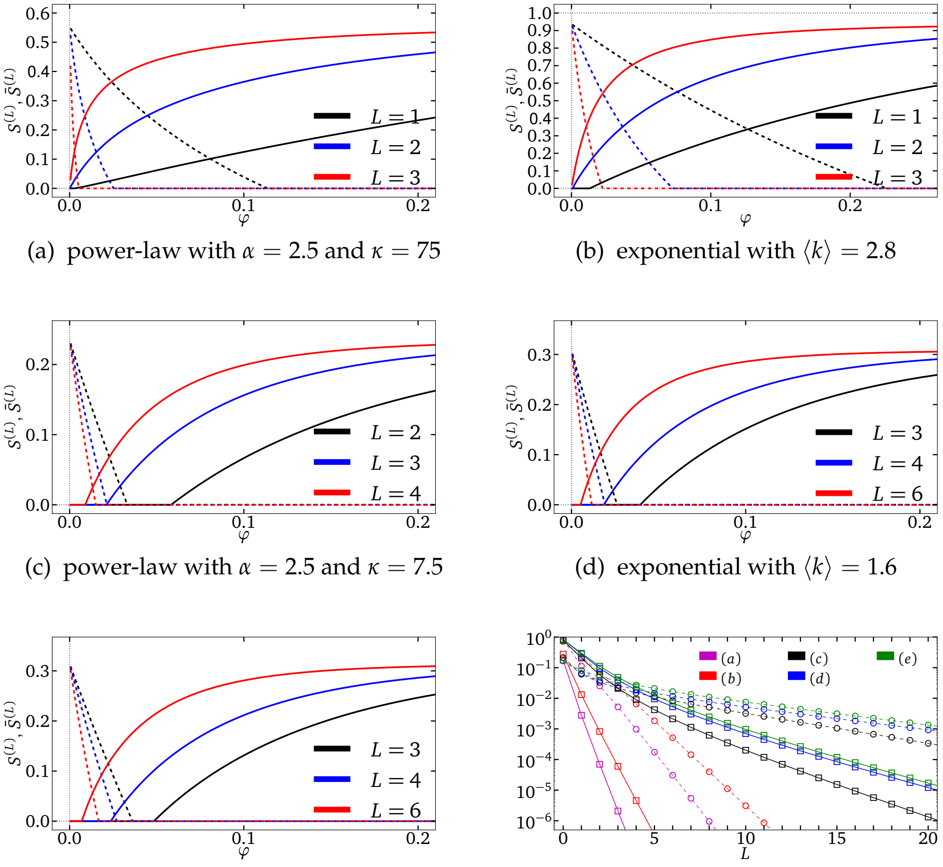
Similarly, the probability $w_i^{(L)}$ for a random node to be of type i is

$$w_i^{(L)} = \begin{cases} \varphi & i=0 \\ (1-\varphi)\big[G_0\big(\chi_{i-1}^{(L)}\big) - G_0\big(\chi_i^{(L)}\big)\big] & 1 \leq i \leq L \\ (1-\varphi)G_0\big(\chi_L^{(L)}\big) & i = L+1 \end{cases} . \tag{5.20}$$

The value of $\varepsilon_0^{(L)}$ and of $w_0^{(L)}$ come from the definition of depth-L percolation itself, that is that type 0 nodes are assigned randomly with probability φ. Using Eqs. (5.18), we see that $\chi_i^{(L)} = (1-\varphi)G_1\big(\chi_{i-1}^{(L)}\big)$ meaning that $\chi_i^{(L)}$ is the probability that the type of the node at the end of a random edge is not lower than i. Hence for $1 \leq i \leq L$, the values of $\varepsilon_i^{(L)}$ and of $w_i^{(L)}$ equal to the probability that the type of the node is not lower than $i-1$ minus the probability that its type is not lower than i. The value of $\varepsilon_{L+1}^{(L)}$ and of $w_{L+1}^{(L)}$ follow directly. Both sets of probabilities are normalized.

We compute the joint degree distribution of each node type in a similar manner. Based on the procedure described above, type i nodes are randomly and independently connected: (i) to no node whose type is lower than $i-1$, (ii) to at least one type $(i-1)$ nodes with probability $\varepsilon_{i-1}^{(L)}$, (iii) to type i nodes with probability $\varepsilon_i^{(L)}$, (iv) to type $(i+1)$ nodes with the complementary probability $\chi_{i+1}^{(L)}$, and (v) and to no node whose type is higher than $i+1$. Enforcing the normalization of the resulting joint degree distributions, we obtain the following associated pgf

$$g_i^{(L)}(\boldsymbol{x}) = \begin{cases} G_0\Big(\varphi x_{0,0} + [1-\varphi]x_{0,1}\Big) & i=0 \\ \dfrac{G_0\Big(\varepsilon_{i-1}^{(L)}x_{i,i-1} + \varepsilon_i^{(L)}x_{i,i} + \chi_{i+1}^{(L)}x_{i,i+1}\Big) - G_0\Big(\varepsilon_i^{(L)}x_{i,i} + \chi_{i+1}^{(L)}x_{i,i+1}\Big)}{G_0\Big(\chi_{i-1}^{(L)}\Big) - G_0\Big(\chi_i^{(L)}\Big)} & 1 \leq i \leq L \\ \dfrac{G_0\Big(\varepsilon_L^{(L)}x_{L+1,L} + \varepsilon_{L+1}^{(L)}x_{L+1,L+1}\Big)}{G_0\Big(\chi_L^{(L)}\Big)} & i = L+1 \end{cases} . \tag{5.21}$$



(a) power-law with $\alpha = 2.5$ and $\kappa = 75$

(b) exponential with $\langle k \rangle = 2.8$

(c) power-law with $\alpha = 2.5$ and $\kappa = 7.5$

(d) exponential with $\langle k \rangle = 1.6$

(e) Poisson with $\langle k \rangle = 1.2$

(f) Thresholds vs. depth

FIGURE 5.3 – Effect of varying the depth L on the coexistence regime. (a)–(e) The size of the non-occupied giant component ($\bar{S}^{(L)}$, dash lines) and of the occupied giant component ($S^{(L)}$, solid lines) are shown as a function of φ for different values of the depth L and different degree distributions. The power-law degree distribution is defined as $P(k) = k^{-\alpha}\mathrm{e}^{-k/\kappa}/\mathrm{Li}_\alpha(\mathrm{e}^{1/\kappa})$ with $k \geq 1$ and $\mathrm{Li}_\alpha(x)$ denoting the polylogarithm. The Poisson degree distribution is defined as $P(k) = \lambda^k \mathrm{e}^{-\lambda}/k!$ with $k \geq 0$ and $\lambda = \langle k \rangle$. See the caption of Fig. 5.2 for the definition of the exponential degree distribution. All curves were obtained by solving Eqs. (5.22)–(5.25). Figures (a)–(e) are a representative subset of the behaviors obtained with many realistic and commonly used degree distributions. (f) Behavior of $\bar{\varphi}_c^{(L)}$ (circles) and $\varphi_c^{(L)}$ (squares) as a function of L using the degree distributions of (a)–(e). Values were obtained from (5.26), and by analyzing the stability of Eqs. (5.23) around $\boldsymbol{a} = \mathbf{1}$. Lines have been added to guide the eye.

Following Ref. [10], we set $x_{L,L+1} = x_{L+1,L} = 1$ in Eqs. (5.21) and the relative size of the giant occupied component, $S^{(L)}$, is computed from

$$S^{(L)} = \sum_{i=0}^{L} w_i^{(L)} \left[1 - g_i^{(L)}(\boldsymbol{a}^{(L)})\right] \tag{5.22}$$

Description	N	$\langle k \rangle$	$k_{\max}$	$G_1'(1)$	$S_{\max}$	Fig.	Ref.
Email communication network of Universitat Rovira i Virgili	1 133	9.08	1 080	125	1 133	5.4b	[83]
Protein interaction network of *S. cerevisiae*	2 640	4.83	111	11.5	2 445	5.4d	[153]
Web of trust of the Pretty Good Privacy (PGP) encryption algorithm	10 680	4.55	205	17.9	10 680	—	[26]
Internet at the level of autonomous systems	22 963	4.22	2 390	260	22 963	5.4c	[4]
arXiv co-authorship network	30 561	8.24	191	20.9	28 502	5.4a	[153]
Gnutella peer-to-peer network	36 682	4.82	55	10.5	36 646	—	[166]
Slashdot online social network	77 360	12.1	2 539	146	77 360	5.4e	[110]
Myspace online social network	100 000	16.8	59 108	3770	100 000	—	[3]
Email exchange network from an undisclosed European research institution	265 009	2.75	7 636	536	224 832	—	[109]
World Wide Web	325 729	6.69	10 721	280	325 729	—	[15]
Polish power grid	3 374	2.41	11	2.15	3 374	5.5b	[201]
Western States Power Grid of the United States	4 941	2.67	19	2.87	4 941	—	[197]
Road network of Pennsylvania	1 088 092	2.83	9	2.20	1 087 562	5.5a	[110]
Road network of Texas	1 379 917	2.79	12	2.15	1 351 137	—	[110]
Road network of California	1 965 206	2.82	12	2.17	1 957 027	—	[110]

TABLE 5.2 – Description and properties of the databases used in Section 5.6 and in Figs. 5.4 and 5.5. The number of nodes (N), the average degree ($\langle k \rangle$), the highest degree ($k_{\max}$), the size of largest connected component ($S_{\max}$) as well as the value of $G_1'(1)$ are given. The databases are divided into two categories: those whose behavior, with regards to observability, is closer to that of a random graph (top), and those whose behavior is similar to that of a lattice (bottom).

[4] Downloaded from `http://www-personal.umich.edu/~mejn/netdata/`.

where $\boldsymbol{a}^{(L)} \equiv \{a_{ij}^{(L)}\}_{i,j=0,\ldots,L}$ is the fixed point—the smallest solution in $[0,1]^{(L+1)^2}$—of the system of equations

$$a_{ij}^{(L)} = \frac{\partial g_j^{(L)}(\boldsymbol{a}^{(L)})/\partial x_{ji}}{\partial g_j^{(L)}(\mathbf{1})/\partial x_{ji}} \tag{5.23}$$

with $i,j = 0,\ldots,L$. The point at which the giant occupied component emerges, $\varphi_c^{(L)}$, is obtained by a linear stability analysis of the fixed point $\{a_{ij}\} = \mathbf{1}$ with $i,j = 0,\ldots,L$. Although the corresponding Jacobian matrix is composed of recurrent patterns of non-zero elements—due to the hierarchy of node types—it has not been possible to extract a useful general equation for $\varphi_c^{(L)}$. The relative size of the non-occupied component, $\bar{S}^{(L)}$, is computed from

$$\bar{S}^{(L)} = w_{L+1}\left[1 - \frac{G_0\left(\varepsilon_L^{(L)} + \varepsilon_{L+1}^{(L)} a_{L+1,L+1}^{(L)}\right)}{G_0\left(\chi_L^{(L)}\right)}\right] , \tag{5.24}$$

where $a_{L+1,L+1}^{(L)}$ is the fixed point of

$$a_{L+1,L+1}^{(L)} = \frac{G_1\left(\varepsilon_L^{(L)} + \varepsilon_{L+1}^{(L)} a_{L+1,L+1}^{(L)}\right)}{G_1\left(\chi_L^{(L)}\right)} . \tag{5.25}$$

Analyzing the stability of the fixed point $a_{L+1,L+1} = 1$, we find that the related critical point, $\bar{\varphi}_c^{(L)}$, is the solution of

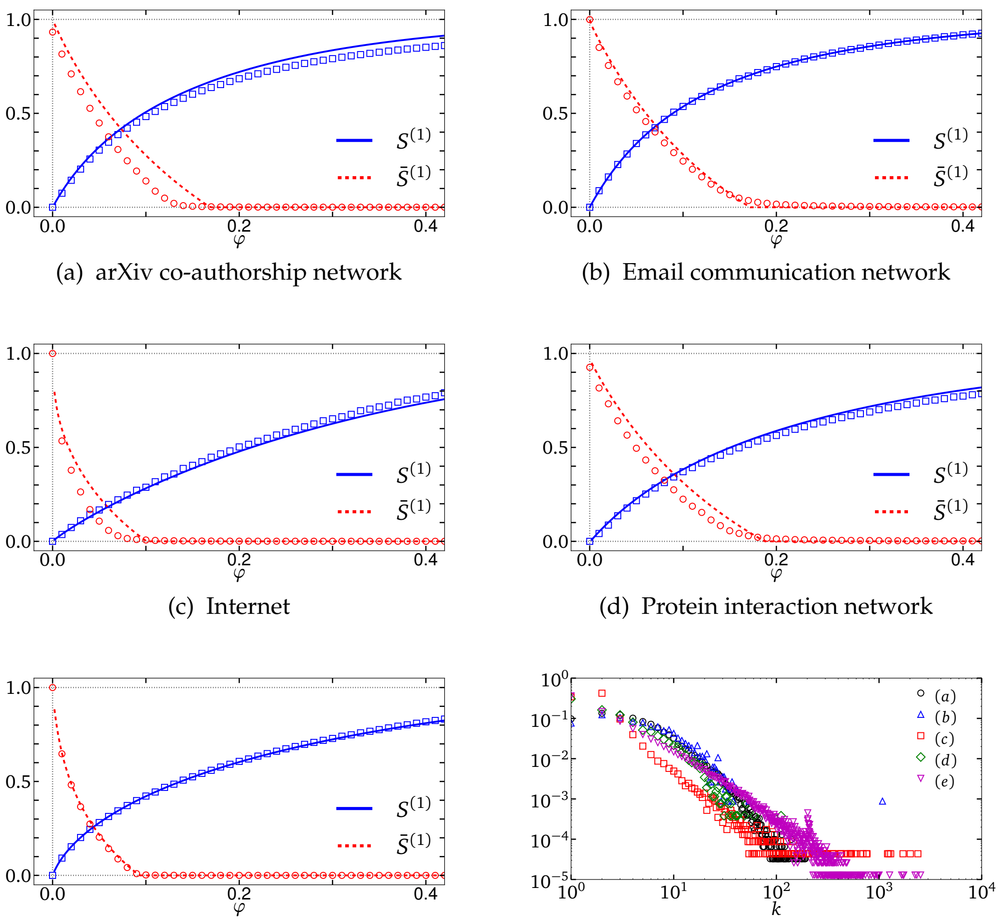
$$(1 - \bar{\varphi}_c^{(L)})G_1'(\chi_L^{(L)}) = 1 . \tag{5.26}$$

Predictions of Eqs. (5.22)–(5.26) are validated in Fig. 5.2c. Equations derived in Sec. 5.4 are retrieved directly by setting $L = 1$ in Eqs. (5.18)–(5.26). A very accurate approximation of Eqs. (5.24) for the case $L = 2$ has been given in the Supplemental Material provided with Ref. [200]. This case is much more delicate than the case $L = 1$: a complete Appendix is devoted to working out the correspondence of the approach of Ref. [200] with the exact calculation provided in this section.

5.5.1 The symmetric case $L = 0$

The case $L = 0$ corresponds to traditional site percolation on random graphs. In the context of observability, it is somewhat trivial as it is symmetric: the non-occupied giant component behaves exactly as the occupied one under the substitution $\varphi \to 1 - \varphi$. It is however an interesting case as expressions for $\varphi_c^{(0)}$ and $\bar{\varphi}_c^{(0)}$ can be obtained in closed form

$$\varphi_c^{(0)} = 1 - \bar{\varphi}_c^{(0)} = \frac{1}{G_1'(1)} . \tag{5.27}$$



(a) arXiv co-authorship network

(b) Email communication network

(c) Internet

(d) Protein interaction network

(e) Slashdot online social network

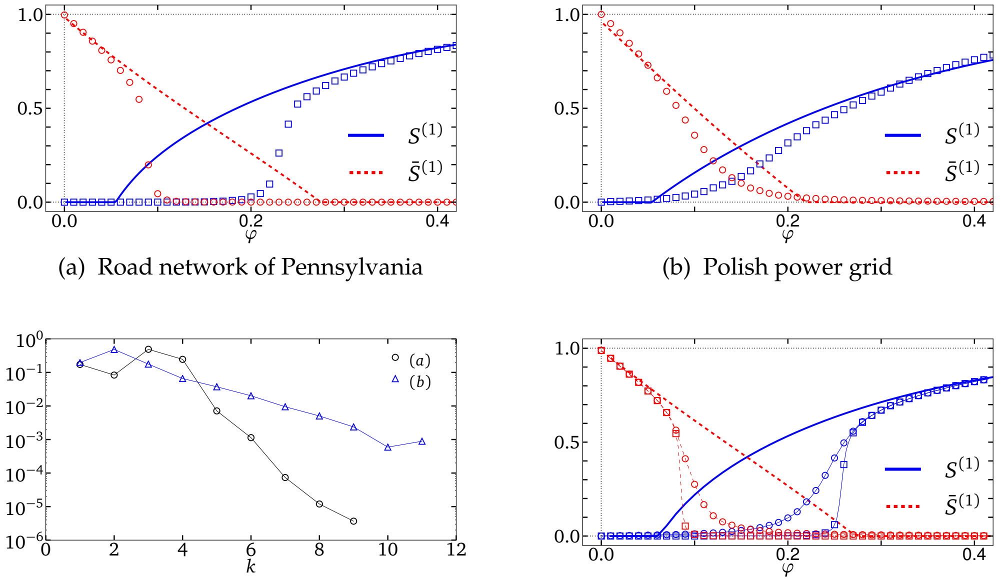
(f) Degree distributions

FIGURE 5.4 – (a)–(e) Depth-1 percolation on graphs extracted from real non-geographically-constrained complex systems (see Table 5.2). Symbols represent the average (100 simulations minimum) relative size of the largest occupied ($S^{(1)}$) and non-occupied ($\bar{S}^{(1)}$) components found in these graphs where directly occupied nodes were selected randomly with probability φ. Lines were obtained by solving Eqs. (5.12), (5.13), (5.15) and (5.16) with the degree distribution extracted from each graph [shown in (f)].

As expected, this corresponds to the threshold value obtained for site percolation on random graphs [37]. Asking for the coexistence of the two extensive components (i.e., $\varphi_c^{(0)} < \bar{\varphi}_c^{(0)}$), we find the condition

$$G_1'(1) > 2\,.$$

This offers a quantitative criterion for the original giant component to be *dense enough* to sustain coexistence: the average excess degree of the original graph ensemble must exceed 2. Recall that in terms of the moments of the degree distribution, $G_1'(1) = (\langle k^2\rangle - \langle k\rangle)/\langle k\rangle$, which permits to rewrite the criterion as $\langle k^2\rangle > 3\langle k\rangle$. As the case $L = 0$ is symmetric under



(a) Road network of Pennsylvania

(b) Polish power grid

(c) Degree distributions

(d) Lattices

FIGURE 5.5 – (a)–(b)Depth-1 percolation on graphs extracted from real geographically-constrained complex systems (see Table 5.2). Symbols represent the average (100 simulations minimum) relative size of the largest occupied ($S^{(1)}$) and non-occupied ($\bar{S}^{(1)}$) components found in these graphs where directly occupied nodes have been selected randomly with probability φ. Lines were obtained by solving Eqs. (5.12), (5.13), (5.15) and (5.16) with the degree distribution extracted from each graph [shown in (c)]. (d) Depth-1 percolation on square $L \times L$ lattices (circles: $L = 70$, squares: $L = 1000$) where edges are randomly removed with probability $p = 0.30$ ($\langle k \rangle = 2.8$). Symbols represent the average (100 simulations minimum) relative size of the largest occupied ($S^{(1)}$) and non-occupied ($\bar{S}^{(1)}$) components. Lines (with no symbols) are the predictions of Eqs. (5.12), (5.13), (5.15) and (5.16) with the binomial degree distribution $P(k) = \binom{4}{k}(1-p)^k p^{4-k}$ with $0 \leq k \leq 4$. Lines have been added between symbols in (c)–(d) to guide the eye.

the substitution $\varphi \to 1 - \varphi$, it is therefore not surprising that coexistence occurs whenever $\varphi_c^{(0)} < 1/2$.

5.5.2 Dependency on the depth L

Using the results of Sec. 5.5, we now investigate the effect of varying L on the coexistence regime. From Eqs. (5.18)–(5.19), it can be shown that for a fixed φ

$$\chi_i^{(L)} = \chi_i^{(L')} > \chi_j^{(L')} \tag{5.28}$$

for $0 \leq i \leq L+1$, $i < j \leq L'+1$ and $L < L'$. This implies that $g_i^{(L)}(\boldsymbol{x}) = g_i^{(L')}(\boldsymbol{x})$ for $0 \leq i \leq L$ and $L < L'$. Because $G_0(x)$ is a monotonous increasing function in [0,1] (as well as its derivatives), we directly see from Eqs. (5.20) that the fraction $w_{L+1}^{(L)}$ of non-occupied nodes decreases with increasing L

$$\frac{w_{L+1}^{(L)}}{w_{L'+1}^{(L')}} = \frac{G_0\big(\chi_L^{(L)}\big)}{G_0\big(\chi_{L'}^{(L')}\big)} = \frac{G_0\big(\chi_L^{(L')}\big)}{G_0\big(\chi_{L'}^{(L')}\big)} > 1$$

for $L < L'$. The more sparse non-occupied nodes are in the graphs, the more likely they will form finite-size components, therefore making an extensive component less likely. Hence we expect $\bar{\varphi}_c^{(L)}$ to decrease with increasing L. In fact, combining Eq. (5.26) and Eq. (5.28) leads to

$$\frac{1-\bar{\varphi}_c^{(L)}}{1-\bar{\varphi}_c^{(L')}} = \frac{G_1'\big(\chi_{L'}^{(L')}\big)}{G_1'\big(\chi_L^{(L)}\big)} = \frac{G_1'\big(\chi_{L'}^{(L')}\big)}{G_1'\big(\chi_L^{(L')}\big)} < 1 \ ,$$

which implies that $\bar{\varphi}_c^{(L)} > \bar{\varphi}_c^{(L')}$ for $L < L'$. The emergence of the giant occupied component is affected in a similar way. As L increases, directly and indirectly occupied nodes represent a larger fraction of the graphs (i.e., $1 - w_{L+1}^{(L)}$) which increases the likelihood of an extensive component. We therefore expect $\varphi_c^{(L)} > \varphi_c^{(L')}$ for $L < L'$. These insights are corroborated by Fig. 5.3. Hence $\varphi_c^{(L)}$ is bounded from above by its value at $L = 0$ [Eq. (5.27)]. This is in accordance with the conclusion of Ref. [200] where it is shown that $\varphi_c^{(1)}$ is bounded from above by a rapidly decreasing function of $G_1'(1)$.

In order to assess the effect of varying the depth L on the coexistence regime, we need to determine how $\varphi_c^{(L)}$ and $\bar{\varphi}_c^{(L)}$ behave relative to each other as L increases. Unfortunately, although the Jacobian matrix determining the stability of the fixed point $\{a_{ij}\} = \mathbf{1}$ looks rather simple [cf. Eq. (5.23)], we have not been able to completely settle this matter analytically. However, as illustrated by Fig. 5.3f, we find that, in all investigated scenarios, $\varphi_c^{(L)}$ decreases faster than $\bar{\varphi}_c^{(L)}$. If this behavior were to be proven true in general, it has the following implications. Firstly, if there is a coexistence interval for a given depth L, then there is a coexistence interval for all $L' > L$. Consequently, increasing the depth L cannot destroy the coexistence regime, it can only bring the bounds of its interval closer to $\varphi = 0$. As a corollary, if $G_1'(1) > 2$, then there exists a coexistence interval for all depth. Note however that although $\varphi_c^{(L)}$ decreases faster than $\bar{\varphi}_c^{(L)}$, the width of the coexistence interval diminishes with increasing L since both threshold values are decreasing [cf. Fig. 5.3f].

Secondly, if no coexistence interval exists for a given depth L, increasing the depth L eventually creates a coexistence regime. This behavior is shown in Figs. 5.3c–(e). Thirdly, the symmetry of the case $L = 0$ implies that both thresholds cannot be greater than 0.5 at the same time for any depth L [see Fig. 5.3f]. As a final remark on the effect of the depth L on the coexistence regime, the fact that $\varphi_c^{(L)}$ appears to be bounded from above by its value at $L = 0$ implies that graphs whose degree distribution's second moment diverges (i.e., *scale-free* degree distributions) would always have a coexistence interval. Indeed, whenever $G_1'(1) \to \infty$, Eq. (5.27) yields $\varphi_c^{(0)} = 0$, and consequently $\varphi_c^{(L)} = 0$ for any L. As *heavy-tailed* degree distribution are ubiquitous in natural and technological complex systems [15, 182, 140, 43, 88, 90], our analysis suggests that coexistence will be found in many real complex systems as $\varphi_c^{(L)}$ will be very close to zero for any depth L.

5.6 Observability of real complex systems

To further our investigation, we simulated depth-1 percolation on graphs representing the underlying web of interactions of real complex systems. A total of 15 systems of diverse nature were considered; details of which are given in Table 5.2. Only a representative subset of our results on those systems are displayed in Figs. 5.4–5.5. Given a random sampling of a fraction φ of the elements of a system (e.g., individuals, autonomous systems, proteins), the mathematical approach introduced in the previous sections allows one to estimate the *coverage* of the system that is achieved given that information about the neighbors up to a distance L of the sampled elements can be gathered as well. This coverage can be estimated in terms of the total number of elements about which information has been obtained (i.e., $\{w_i^{(L)}\}_{i=0,\ldots,L+1}$), or in terms of the largest number of contiguous elements (i.e., $S^{(L)}$), as in the main focus of this work.

Two examples will serve to explain the practical utility of our approach. Suppose that we want to get a global picture of the scientists working in a specific field without any prior information about that field. One way to achieve this is to browse the latest table of contents of appropriate journals, to identify scientists that have published something relevant to that field and then find with whom they co-authored papers during their careers. Although a sampling through the table of contents is not rigorously equivalent to the random and uncorrelated sampling considered in the previous sections, the quality of the coverage obtained can be estimated by studying depth-1 percolation on the associated co-authorship network. Looking up the co-authors of these co-authors up to a distance L then corresponds to depth-L percolation. Similarly, it has recently been revealed that intelligence agencies may gather information on individuals that are up to three "hops" (i.e., $L = 3$) from suspected individuals [40]. Again, our model offers a theoretical framework to estimate the extent of the population that could be investigated by studying the depth-L percolation of online social networks, email communications or mobile phone networks.

Figures 5.4–5.5 summarize the typical behaviors obtained when simulating depth-1 percolation on the graphs described in Table 5.2. Our results suggest that real systems behave differently with regards to observability according to whether they are geographically-constrained or not. Graphs that are not geographically-constrained behave more or less like random graphs (long-range connections are allowed), while geographically-constrained graphs behave more like lattices (no long-range connections).

We find that the observability of non-geographically-constrained graphs [Figs. 5.4a–(e)] is surprisingly well predicted by our mathematical framework despite the fact that most of these graphs have a far less trivial structure (e.g., clustering, correlations) than the Configuration Model which considers graphs that are random in all aspects other than the degree distribution. More importantly, we determine that these graphs have a structure that permits a coexistence regime. These graphs also display a vanishing threshold, $\varphi_c^{(1)}$, for the observable giant component. This agrees with the prediction of our model since these graphs have very skewed (i.e., scale-free) degree distribution [see Fig. 5.4f].

Contrariwise, our results for geographically-constrained graphs [Figs. 5.5a–(b)] display totally different behaviors[5]. Apart from the non-zero threshold for the occupied giant component, $\varphi_c^{(1)}$, caused by their approximatively exponential degree distributions [see Fig. 5.5c], the behavior of the two extensive components is poorly predicted by our mathematical framework. A fairly large coexistence interval is predicted while numerical simulations show that their structure does not, or barely, allow for a coexistence regime. Geographically-constrained graphs seem to be more accurately modeled by lattices than by random graphs. We have simulated depth-1 percolation on $L \times L$ square lattices where a fraction p of edges are randomly removed. As shown in Fig. 5.5d, by simply choosing p to match their average degree and L to match their size, we have been able to qualitatively reproduce the results obtained with the real graphs [i.e., Fig. 5.5a–(b)]. Although preliminary, these results point towards the topological properties that should be incorporated in a future theoretical formalism to accurately model geographically-constrained graphs.

5.7 Concluding remarks

We have presented a general theoretical framework to study the observability of random graphs. On the one hand, it has allowed us to demonstrate that two extensive components, an observable and non-observable, may coexist for a wide range of realistic parameters, and that coexistence can be observed in many real complex systems. Our results suggest that coexistence could be an impediment to the monitoring of large real systems, and should therefore be considered in future investigations. On the other hand, the mapping of depth-L percolation unto multitype graphs opens the way to the use of recent developments in percolation theory to study graphs with more realistic structures (e.g., clustering, correlations), and to investigate the efficiency of various distribution schemes for the monitoring units (e.g., according to the degree, to the local clustering or to the centrality of nodes) [6, 8, 92, 86, 46, 178]. We have also shown that our approach performs poorly at predicting the observability of geographically-constrained systems, and that achieving large-scale observability of these systems requires more monitoring units than suggested by calculations based on the Configuration Model. We have provided numerical evidences that these systems in fact behave more like lattices than random graphs. This observation raises many

[5] This conclusion could also be drawn by comparing the black curve of Fig. 2 with the solid gray curve of Fig. 3(a) in Ref. [200].

questions whose answers are expected to improve our understanding of the organization of these complex systems, and consequently to improve our capability to predict their behavior.

5.8 Appendix: Comparison with the approach of Yang *et al.* for $L = 2$

In this section we analyze the solution to the observability problem given by Yang *et al.* in the case $L = 2$ and compare it with the prediction of our formalism.

5.8.1 Multitype formalism

Let us first explicit the predictions of our approach. To lighten the presentation, we omit the superscript specifying the depth since this entire section focuses on the case $L = 2$. Setting $L = 2$ in Eq. (5.18), we obtain

$$\begin{aligned}
\varepsilon_0 &= \varphi && \text{(5.29a)}\\
\varepsilon_1 &= (1-\varphi)\big[1 - G_1(1-\varphi)\big] && \text{(5.29b)}\\
\varepsilon_2 &= (1-\varphi)\big[G_1(1-\varphi) \\
&\quad - G_1\big((1-\varphi)G_1(1-\varphi)\big)\big] \,, && \text{(5.29c)}
\end{aligned}$$

where we have omitted the case $i = 3$ since the present section focuses on the giant observable component solely. Similarly, Eq. (5.19) becomes

$$\begin{aligned}
\chi_0 &= 1 && \text{(5.30a)}\\
\chi_1 &= 1-\varphi && \text{(5.30b)}\\
\chi_2 &= (1-\varphi)G_1(1-\varphi) && \text{(5.30c)}\\
\chi_3 &= (1-\varphi)G_1\big((1-\varphi)G_1(1-\varphi)\big) \,, && \text{(5.30d)}
\end{aligned}$$

and Eq. (5.20) yields

$$\begin{aligned}
w_0 &= \varphi && \text{(5.31a)}\\
w_1 &= (1-\varphi)\big[1 - G_0(1-\varphi)\big] && \text{(5.31b)}\\
w_2 &= (1-\varphi)\big[G_0(1-\varphi) \\
&\quad - G_0\big((1-\varphi)G_1(1-\varphi)\big)\big] \,. && \text{(5.31c)}
\end{aligned}$$

Combining Eqs. (5.29)–(5.30) with Eqs. (5.21) and (5.23), we obtain the following system of equations

$$a_{00} = G_1\big(\varphi a_{00} + (1-\varphi)a_{01}\big) \tag{5.32a}$$

$$a_{01} = G_1(\varphi a_{00} + \varepsilon_1 a_{11} + \chi_2 a_{12}) \tag{5.32b}$$

$$a_{11} = \frac{G_1(\varphi a_{00} + \varepsilon_1 a_{11} + \chi_2 a_{12}) - G_1(\varepsilon_1 a_{11} + \chi_2 a_{12})}{1 - G_1(1-\varphi)} \tag{5.32c}$$

$$a_{12} = \frac{G_1(\varepsilon_1 a_{11} + \varepsilon_2 a_{22} + \chi_3)}{G_1(1-\varphi)} \tag{5.32d}$$

$$a_{22} = \frac{G_1(\varepsilon_1 a_{11} + \varepsilon_2 a_{22} + \chi_3) - G_1(\varepsilon_2 a_{22} + \chi_3)}{G_1(1-\varphi) - G_1\big((1-\varphi)G_1(1-\varphi)\big)} \tag{5.32e}$$

whose fixed point determines the size and behavior of the giant observable component. As for the case $L = 1$, some a_{ij} are equal: $a_{10} = a_{00}$ and $a_{21} = a_{11}$ when $L = 2$. In fact, since the directly observable nodes (type 0) are randomly distributed and the type of the other nodes is inherited by the type of their neighbors, we find in general that $a_{i+1,i} = a_{i,i}$. In other words, the excess degree distribution of node A is independent of the type of the node from which it has been reached, as long as it is not the type of this neighbor that defines the type of node A. Note however that we will use interchangeably $a_{i+1,i}$ and $a_{i,i}$ according to whether it simplifies the notation or clarifies the significance of mathematical quantities. Combining Eqs. (5.31) with Eqs. (5.21)–(5.22), the size of the giant observable component is given by

$$\begin{aligned} S = 1 &- \varphi G_0\big(\varphi a_{00} + (1-\varphi)a_{01}\big) \\ &- (1-\varphi)\Big\{G_0(\varphi a_{00} + \varepsilon_1 a_{11} + \chi_2 a_{12}) \\ &- G_0(\varepsilon_1 a_{11} + \chi_2 a_{12}) + G_0\big((1-\varphi)G_1(1-\varphi)\big) \\ &+ G_0(\varepsilon_1 a_{11} + \varepsilon_2 a_{22} + \chi_3) - G_0(\varepsilon_2 a_{22} + \chi_3)\Big\}\,. \end{aligned} \tag{5.33}$$

5.8.2 Yang *et al.*'s approach

Let us now recall the equations for $L = 2$ as given in the Supplemental Material of Ref. [200]. The authors define three probabilities u, v and s which are analogous to the $\{a_{ij}\}$ used in our approach: they correspond to the probability that a given randomly chosen edge does not lead to the giant observable component. These probabilities are defined as follows. (i) u is the probability that an edge stemming from a node of type 0 (i.e., directly observable) does not lead to the giant observable component. (ii) v is the probability that an edge stemming from a node of type 1 towards a node of type 1 or of type 2 does not lead to the giant observable component. (iii) s is the probability that an edge stemming from a node of type 2 does not lead to the giant observable component. The authors then explain that by following a similar argument to the one used for $L = 1$, it can be shown that

$$\begin{aligned} u &= \varphi G_1(u) + (1-\varphi) G_1(\psi_1) && \text{(5.34a)} \\ v &= G_1\big((1-\varphi)s\big) + \psi_2 && \text{(5.34b)} \\ s &= G_1\big((1-\varphi)G_1(1-\varphi)\big) + \psi_2 \\ &\quad + G_1(\psi_3) - G_1\big((1-\varphi)G_1(1-\varphi)s\big) && \text{(5.34c)} \end{aligned}$$

where

$$\begin{aligned} \psi_1 &= \varphi G_1(u) + (1-\varphi)v && \text{(5.34d)} \\ \psi_2 &= G_1(\psi_1) - G_1\big((1-\varphi)v\big) && \text{(5.34e)} \\ \psi_3 &= (1-\varphi)\psi_2 + (1-\varphi)G_1(1-\varphi)s\ , && \text{(5.34f)} \end{aligned}$$

and that the size of the giant observable component is given by

$$\begin{aligned} S_Y = 1 - \varphi G_0(u) - (1-\varphi)\Big\{ & G_0(\psi_1) - G_0\big((1-\varphi)v\big)\big) \\ & + G_0(\psi_3) - G_0\big((1-\varphi)G_1(1-\varphi)s\big)\Big\} \\ & + G_0\big((1-\varphi)G_1(1-\varphi)\big)\ . \end{aligned} \tag{5.35}$$

5.8.3 Comparison of the two approaches

We now investigate whether these two approaches are equivalent or not. As mentioned above, u is the probability that a directly observable node (type 0) is not linked to the giant observable component via one specific edge. This corresponds to the probability that the node at the other end of the edge, say node B, is of type 0 (probability ε_0) and that the edge does not lead to the giant observable component (probability a_{00}), or that node B is of type 1 (probability χ_1) and that the edge does not lead to the giant observable component (probability a_{01}). Summing these two contributions and then using Eqs. (5.32), we find

$$\begin{aligned} u &= \varepsilon_0 a_{00} + \chi_1 a_{01} \\ &= \varphi G_1\big(\underbrace{\varphi a_{00} + (1-\varphi)a_{01}}_{u}\big) \\ &\quad + (1-\varphi)G_1(\underbrace{\varphi a_{10} + \varepsilon_1 a_{11} + \chi_2 a_{12}}_{\psi_1})\ , \end{aligned} \tag{5.36}$$

which corresponds to Eq. (5.34a) provided that the identification of ψ_1 holds. As in the case $L = 1$, ψ_1 is the probability that a node of type 1 is not connected to the giant observable component via a specific edge. Three different scenarios must be accounted for depending on the type of the node at the other end of the edge: this node can be of type 0, type 1 or type 2, with probability ε_0, ε_1 and χ_2, respectively. Multiplying each probability by the corresponding probability that the edge does not lead to the giant component, we retrieve the above identification

$$\begin{aligned} \psi_1 &= \varepsilon_0 a_{10} + \varepsilon_1 a_{11} + \chi_2 a_{12} \\ &= \underbrace{\varphi G_1\big(\varphi a_{00} + (1-\varphi)a_{01}\big)}_{\varphi G_1(u)} + \underbrace{\varepsilon_1 a_{11} + \chi_2 a_{12}}_{(1-\varphi)v} \, , \end{aligned} \tag{5.37}$$

which corresponds to Eq. (5.34d) provided that the identification of v holds. The first term on the right-hand side of this last equation corresponds to the situation where the node at the other end of the edge is of type 0 (probability φ) and does not lead to the giant observable component [probability $G_1(u)$]. The second term corresponds to the case where the neighboring node is of type 1 or of type 2, which occurs with probability $1 - \varphi$ (recall that the neighbor of a node of type 1 cannot be of type 3, by definition), and that this edge does not lead to the giant component, which by definition occurs with probability v. In terms of the formalism that we propose, the probability for an edge leaving a node of type 1 to lead to a node of type 1 is ε_1, and is χ_2 if the neighboring node is of type 2 instead. Weighting these probabilities with the appropriate probability that the edge does not lead to the giant observable component yields precisely

$$\begin{aligned} (1-\varphi)v &= \varepsilon_1 a_{11} + \chi_2 a_{12} \\ &= (1-\varphi)\Big\{ \underbrace{\underbrace{G_1(\varphi a_{10} + \varepsilon_1 a_{11} + \chi_2 a_{12})}_{G_1(\psi_1)} - G_1(\underbrace{\varepsilon_1 a_{11} + \chi_2 a_{12}}_{(1-\varphi)v})}_{\psi_2} + G_1(\underbrace{\varepsilon_1 a_{21} + \varepsilon_2 a_{22} + \chi_3}_{(1-\varphi)s}) \Big\} \, , \end{aligned} \tag{5.38}$$

from which we retrieve Eq. (5.34e) and Eq. (5.34b) provided that the identification of s holds. We see from this last equation that ψ_2 is the probability that a node of type 1 reached from a node of type 1 does not lead to the giant observable component. Note that because it has been reached from a node of type 1, this node must have at least one neighbor of type 0 in order to be of type 1. Since $G_1(\psi_1)$ includes the case where all neighbors of a node of type 1 are not of type 0, the probability of such an event must be removed from the count, which is achieved by subtracting $G_1\big((1-\varphi)v\big)$. Additionally if the node at the other end of the edge is of type 2 instead of type 1, then none of its *other* neighbors must be of type 0, which occurs individually with probability $1 - \varphi$, and must not lead to the giant component, which by definition occurs with probability s. Averaging over the number of *other* neighbors [the excess degree distribution generated by $G_1(x)$], we obtain the third term on the right-hand side of Eq. (5.38). Again, the probability $(1-\varphi)s$ can be expressed in terms of our formalism. The probability that an edge leaving a node of type 2 towards a node of type 1, of type 2 and of type 3 is respectively ε_1, ε_2 and χ_3. Weighting ε_1 and ε_2 by the probability that the edge does not lead to the giant component (recall that a node of type 3 does not belong to the giant observable component "with probability 1") yields our previous identification

$$\begin{aligned}(1-\varphi)s &= \varepsilon_1 a_{21} + \varepsilon_2 a_{22} + \chi_3 \\ &= (1-\varphi)\Big\{\psi_2 + \big[G_1\big(\underbrace{(1-\varphi)\psi_2 + \varepsilon_2 a_{22} + \chi_3}_{\psi_3}\big) - G_1(\varepsilon_2 a_{22} + \chi_3)\big] \\ &\qquad\qquad + G_1\big((1-\varphi)G_1(1-\varphi)\big)\Big\} \ , \end{aligned} \tag{5.39}$$

where we have used the fact that $\varepsilon_1 a_{21} = \varepsilon_1 a_{11} = (1-\varphi)\psi_2$ and the definition of χ_3. Comparing this last equation with Eqs. (5.34c) and (5.34f), we find that the two approaches are equivalent if

$$(1-\varphi)G_1(1-\varphi)s = \varepsilon_2 a_{22} + \chi_3 \ . \tag{5.40}$$

As for ψ_2, we see from Eq. (5.39) that ψ_3 is the probability that an edge between two nodes of type 2 does not lead to the giant observable component. Since the node of type 2 reached from such edge does not inherit its type from the node of type 2 at the other end of the edge, at least one of its *other* neighbors must be of type 1. Again, since $G_1(\psi_3)$ includes the configuration where every *other* neighbors of the node of type 2 are of type 2 or of type 3, this eventuality must be removed from the count, which is achieved by subtracting $G_1(\varepsilon_2 a_{22} + \chi_3)$.

Let us now investigate the validity of Eq. (5.40). Replacing $(1-\varphi)s$ by $\varepsilon_1 a_{21} + \varepsilon_2 a_{22} + \chi_3$ yields the following alternative criterion for the complete equivalence of the two approaches

$$\varepsilon_2 a_{22} + \chi_3 = (\varepsilon_1 a_{21} + \varepsilon_2 a_{22} + \chi_3)G_1(1-\varphi) \ , \tag{5.41}$$

which is most certainly not true in general. Although very similar, these two approaches are therefore not strictly equivalent. In fact, their numerical predictions differ by less than a fraction of one percent in all investigated cases. This difference stems for the use of s for two different purposes in the approach presented in Ref. [200]. On the one hand, s is initially defined as the probability that an edge stemming out of a node of type 2 does not lead to the giant observable component irrespective of the type of the node at its other end (i.e., type 1, type 2 or type 3). On the other hand, as it is used in Eqs. (5.34c) and (5.34f), the possibility that the type of the node at the other end is of type 1 is excluded since it is taken care of by the probability $(1-\varphi)\psi_2$. More precisely, in our formalism $(1-\varphi)G_1(1-\varphi) = \varepsilon_2 + \chi_3$ is the probability that the node at the other end of an edge and its *other* neighbors are not of type 0. Since this edge is leaving a node of type 2, the node at its other end is of type 2 or of type 3; it therefore cannot be of type 1. In other words, the term $(1-\varphi)G_1(1-\varphi)s$ uses s as an approximation of the probability that an edge leaving a node of type 2 towards a node of type 2 or type 3 does not lead to the giant observable component.

Conclusion et perspectives

L'étude des systèmes complexes est l'un des principaux défis de la science au XXI^{e} siècle, et la physique théorique, riche de ses concepts et de ses techniques, est avantageusement positionnée pour y contribuer significativement. Nous avons montré dans cette thèse comment la physique statistique, la dynamique non linéaire et la théorie de la percolation pouvaient être mises à profit pour décrire la structure de graphes et, ce faisant, contribuer à élucider la relation fondamentale entre les interactions au sein de systèmes complexes et leurs propriétés macroscopiques émergentes.

Nous avons présenté un formalisme permettant d'obtenir de façon exacte plusieurs propriétés structurelles d'un ensemble général de graphes aléatoires [ex. taille et probabilité d'une composante extensive ($\mathcal{S}$ et $\mathcal{P}$), seuil de percolation ($\det(\mathbf{J} - \mathbf{I}) = 0$), et tailles des petites composantes ($K(z)$)]. En adoptant une approche multitype, nous avons, d'une part, fait de cet ensemble une synthèse des modèles publiés au cours des quinze dernières années, et, d'autre part, proposé le modèle de percolation sur graphes aléatoires le plus général à ce jour, améliorant grandement la complexité des graphes pouvant être considérés dans un formalisme analytique exact.

La polyvalence de notre formalisme lui confère également le rôle de *laboratoire théorique*. En effet, sa flexibilité lui permet de reproduire avec précision un large spectre de propriétés structurelles, et ainsi de pouvoir comparer leur effet respectif sur le même pied. Il est donc possible d'utiliser notre formalisme pour valider et approfondir différentes hypothèses de recherche. À cet égard, le projet présenté à l'annexe A est un exemple probant.

L'adaptation de notre formalisme à la modélisation de systèmes *interdépendants* illustre également la polyvalence de notre approche et ouvre la voie à plusieurs avenues de recherche. Étonnamment, depuis la démonstration d'une transition de phase discontinue, les modèles incorporant des propriétés structurelles plus réalistes aux graphes interdépendants n'apparaissent qu'au compte-goutte. Notre contribution fait donc d'une pierre deux coups : nous introduisons le modèle de percolation sur graphes aléatoires le plus général qui soit et transposons ses capacités directement à la modélisation de graphes interdépendants. Les résultats préliminaires suggérant une relation entre l'effet d'agrégation et l'amplitude de la discontinuité à la transition de phase illustrent bien le type de contributions que permet notre approche.

Dans un esprit similaire, un intérêt toujours grandissant à l'égard des graphes *multiplex*, dans lesquels les interactions s'effectuent à plusieurs niveaux, s'est développé au cours des dernières années [25, 48, 52, 105, 107]. Par exemple, les villes d'un pays sont connectées les unes aux autres via les réseaux routier, ferroviaire et aérien. Cette généralisation requiert la redéfinition de plusieurs concepts, tels que l'effet d'agrégation (ex. quelle importance donne-t-on à un triangle formé de liens appartenant à différents niveaux ?), et, par le fait même, nécessite l'adaptation des modèles développés pour étudier les graphes *monoplex* (c.-à-d. un seul niveau d'interaction). Étant donnée la présence de types de demi-liens, notre modèle permet déjà la modélisation de graphes *multiplex* : ils ne sont qu'un cas particulier de l'ensemble général de graphes aléatoires que nous avons développé. Nos travaux offrent donc un outil théorique riche et polyvalent pour quiconque s'intéresse aux graphes *multiplex*.

La comparaison des prédictions de notre formalisme avec les résultats obtenus numériquement sur des graphes extraits de systèmes complexes réels suggère toutefois qu'il y a encore place à amélioration (voir le chapitre 5 et l'annexe A). En effet, malgré l'ensemble des propriétés réalistes pouvant désormais être incluses dans un modèle théorique, nous ne sommes toujours pas en mesure de correctement reproduire, de façon effective, la structure des graphes réels. Bien qu'il ne soit pas certain qu'une telle représentation effective soit possible ou que celle-ci soit universelle, les résultats présentés dans cette thèse suggèrent deux pistes à explorer.

Une première piste concerne l'effet d'agrégation. Les quantités scalaires avec lesquelles nous mesurons cet effet ne révèlent que partiellement la façon dont l'agrégation s'organise dans un graphe. Par conséquent, bien que nous soyons en mesure d'inclure cet effet dans notre formalisme, et de générer des graphes reproduisant les coefficients d'agrégation mesurés pour des graphes réels, rien ne nous assure que l'agrégation ait été fidèlement reproduite. Quelques nouvelles façons de quantifier l'agrégation ont été proposées, mais aucune d'entre elles ne s'est encore démarquée du lot [46, 45, 177, 202]. Il est donc important de poursuivre les recherches en ce sens, puisqu'une façon systématique de quantifier l'organisation de l'agrégation dans un graphe pourrait alors être intégrée dans un formalisme théorique, et ultimement mener à de meilleurs prédictions.

Une deuxième piste provient des résultats obtenus à l'aide des graphes extraits des réseaux routiers et de distribution électrique. De par leur nature, ces graphes vivent dans des espaces de faible dimension (c.-à-d. 2 ou 3 dimensions). Ceci contraint fortement leur structure notamment en restreignant les voisins potentiels d'un nœud à son voisinage géographique immédiat. Cet effet n'est pas reproduit d'emblée par les modèles de graphes aléatoires. Les conclusions tirées à l'annexe A suggèrent toutefois qu'il soit possible de convertir ces contraintes en corrélations entre des types de nœuds différents. Ceci ouvre la porte à l'éventuelle inclusion d'une structure géographique effective dans un modèle de graphes aléatoires. Quelques modèles de graphes contraints géographiquement ont déjà été proposés, mais ces modèles, essentiellement numériques, ne permettent qu'un traitement analytique limité et ont, de ce fait, un faible pouvoir prédictif [17, 18, 70, 171]. Néanmoins, une analyse en profondeur de ces modèles pourrait permettre d'identifier l'élément-clé qui mènera à l'ajout de ces contraintes géographiques dans un formalisme futur.

Les systèmes complexes sont un thème de recherche méconnu de la physique. Ils sont pourtant des exemples probants de la polyvalence de la physique théorique et, par conséquent, de la pertinence du rôle des théoriciens dans des domaines s'écartant des sujets de recherche traditionnels. En les dépouillant du superflu, ces systèmes complexes peuvent être ramenés à l'essentiel, ouvrant ainsi la voie à l'élaboration de modèles théoriques permettant d'élucider le lien entre leur organisation et leur dynamique. Alors que l'on ne pourrait y voir qu'une simple digression, nous soutenons que ce nouveau thème de recherche est une suite logique aux efforts déployés en physique depuis des siècles. En effet, étant donnée l'omniprésence des systèmes complexes dans notre environnement, la recherche théorique dont ils sont l'objet permet de repousser les limites de notre compréhension du monde qui nous entoure, et il s'agit là de l'essence même de la quête du physicien.

Annexe A

Application II : Graphes aléatoires avec structure *k-core* quelconque

Article original :

Random networks with arbitrary k-core structure

Laurent Hébert-Dufresne [1], **Antoine Allard**[1], Jean-Gabriel Young et Louis J. Dubé

Département de Physique, de Génie Physique, et d'Optique,
Université Laval, Québec (Qc), Canada G1V 0A6

Référence : Phys. Rev. E, **88** 062820 (2013) [92]

© 2013 American Physical Society (§A.2 –§A.9) [2]

[1] Le développement théorique, les simulations numériques de même que la rédaction du manuscrit sont essentiellement les fruits du travail des deux premiers auteurs. Les autres auteurs ont participé à l'élaboration générale du projet de même qu'à la révision du manuscrit.

[2] Ces sections contiennent le contenu original de l'article. Celui-ci n'a été modifié que pour se conformer au format exigé par la Faculté des études supérieures et postdoctorales de l'Université Laval.

A.1 Avant-propos

Cet article présente le fruit d'une collaboration étroite avec Laurent Hébert-Dufresne et est un bel exemple de l'utilisation visée du modèle général présenté au chapitre 4, c.-à-d. celui de *laboratoire théorique*. Ce projet fut entrepris pour répondre à la question suivante : Dans quelle mesure la structure en couches (en anglais *k-core decomposition*, à être définie sous peu) peut-elle agir comme structure effective des graphes extraits de systèmes complexes réels ? Initié à l'origine par Laurent Hébert-Dufresne, ce projet n'était que de nature numérique : nous ne cherchions alors qu'une façon de recréer des graphes possédant une structure en couches quelconque.

Ce n'est qu'à partir du moment qu'une recette fut établie et que les résultats préliminaires furent prometteurs que nous nous sommes intéressés à une description mathématique des graphes ainsi générés. À cet égard, le modèle présenté au chapitre 4 nous offrait le cadre théorique idéal. D'une part, nous avons été en mesure de reproduire les résultats que nous nous ne générions alors que par le biais de simulations numériques coûteuses en ressources informatiques. Dorénavant ceux-ci s'obtenaient en quelques secondes/minutes en solutionnant des équations analogues à l'équation (1.64). D'autre part, la perspective que nous offrit la description mathématique nous permit de raffermir et de simplifier la recette avec laquelle nous générions les graphes (ex. l'algorithme Metropolis-Hastings introduit à la section A.5.3).

Nos résultats nous permirent de conclure que la structure en couches n'est pas suffisante pour décrire effectivement la structure d'un graphe, mais que cette description est au moins aussi précise que d'autres mesures déjà établies. Qui plus est, le modèle général nous permit de démontrer que cette approche était plus efficace autant du point de vue computationnel que du point de vue informationnel.

Quantity	Description	Definition
N	Number of nodes	Sec. A.4
T	Probability for an edge to not be removed during the percolation process	Sec. A.4
$P(k)$	Degree distribution	Sec. A.4
$P(k,k')$	Joint degree distribution	Sec. A.4
$k_{\max}$	Highest degree	Sec. A.4
$c_{\max}$	Highest coreness	Sec. A.5.2
$\mathbf{K}$	Matrix whose elements K_{ck} correspond to the fraction of the nodes that have a coreness c and a degree k	Sec. A.5.2
$\mathbf{C}$	Matrix whose elements $C_{cc'}$ give the fraction of edges that leave nodes of coreness c to nodes of coreness c'	Sec. A.5.2
w_c	Fraction of nodes whose coreness is equal to c	Eq. (A.1)
$P_c(\boldsymbol{k})$	Joint degree distribution of nodes of coreness c	Eq. (A.2)
$\langle k \rangle_c$	Average degree of nodes of coreness c	Eq. (A.3)
$\langle k \rangle$	Average degree of nodes	Eq. (A.4)
$R(c', j\|c, i)$	Probability that a node of coreness c through a stub of color i leads to a node of coreness c' through one of its stubs of color j	Sec. A.5.2
$g_c(\boldsymbol{x})$	pgf generating the joint degree distribution of nodes of coreness c	Eq. (A.6)
$f_{ci}(\boldsymbol{x})$	pgf generating the joint excess degree distribution of nodes of coreness c that have been reached through their stub of color i	Eqs. (A.7) and (A.8)
S	Size of the giant component	Eq. (A.9)
a_{ci}	Probability that a node of coreness c reached via one of its stub of color i does not lead to the giant component	Eq. (A.10)
$\Gamma(c, i; c', j)$	Wanted fraction of edges that join nodes of coreness c and c' via their respective stubs of color i and j	Eqs. (A.11)
C	Clustering coefficient	Sec. A.5.4
r	Degree correlation coefficient	Sec. A.5.4
ℓ	Mean shortest path	Sec. A.5.4

TABLE A.1 – Glossary of the major mathematical objects defined in this chapter.

A.2 Abstract

The k-core decomposition of a network has thus far mainly served as a powerful tool for the empirical study of complex networks. We now propose its explicit integration in a theoretical model. We introduce a Hard-core Random Network model that generates maximally random networks with arbitrary degree distribution *and* arbitrary k-core structure. We then solve exactly the bond percolation problem on the HRN model and produce fast and precise analytical estimates for the corresponding real networks. Extensive comparison with real databases reveals that our approach performs better than existing models, while requiring less input information.

A.3 Introduction

We address the challenge of designing a realistic model of complex networks while preserving its analytic tractability. The model should include the essential structural properties of real networks, and the theoretical framework should guarantee easy access to quantitative calculations. For the second aspect of this endeavour, we cast our analysis in terms of a percolation problem. This has been a topic of choice for some years since it can just as well represent the dynamics *of* a network as the dynamics *on* the network [13, 44, 58, 59, 91, 94, 124, 144]. One might think of its growth, its robustness (to attacks or failures) and the propagation of emerging infectious agents (e.g. disease or information).

While the study of percolation models on idealized networks has led to a better understanding of both the processes they model and the networks that support them, the study of percolation on real networks has somewhat stagnated. Unfortunately, purely numerical approaches are time-consuming, require a complete description of the networks under scrutiny and lack the insights of an analytical description. Conversely, although analytical modeling provides a better understanding of the organization of real networks, they are limited at present to simplified random models (see [139, 144] and references therein).

In this paper, we demonstrate how the k-core structure of networks (hereafter simply core structure) plays a central role in the outcome of bond percolation, and how it acts as a proxy that captures the essential structural properties of real networks. The ensuing model, that we call the Hard-core Random Network (HRN) model, creates maximally random networks with an arbitrary degree distribution *and* an arbitrary core structure. We also propose a Metropolis-Hastings algorithm to generate such random networks. The HRN model serves our purpose well since it is shown to be amenable to an exact solution for the size of the extensive "giant" component (in the limit of large network size). With less input information, it outperforms the current standard model [122] for precise prediction of percolation results on real networks.

The organization of this paper goes as follows. In Sec. A.4, we introduce the bond percolation problem and briefly present the two models used for comparison. In Sec. A.5, we present the HRN model, the equations used to solve the bond percolation problem and the Metropolis-Hastings algorithm generating the corresponding random networks. We also compare the predictions of the HRN model and the ones of the two aforementioned models with the results obtained numerically using real network databases. Final remarks are collected in the last section.

A.4 Bond percolation on networks

The bond percolation problem concerns the connectivity of a network after the removal of a fraction $1 - T$ of its edges. More precisely, for a synthetic or empirical network, we are interested in the fraction S of nodes contained in the largest connected component—the giant component—after each edge has been removed independently with a probability $1 - T$. In the limit of large networks, this component undergoes a *phase transition* at a critical point T_c during which its size (the number of nodes it contains) becomes an extensive quantity that scales linearly with the number of nodes (N) of the whole network [42].

To compare and assert the precision of the predictions of our model, we use the *Configuration Model* (CM) and *Correlated Configuration Model* (CCM) [135, 136, 147, 193] as benchmarks [see Fig. A.1a–(b)]. These models define maximally random network ensembles that are random in all respects other than the degree distribution (CM,CCM) and the degree-degree correlations (CCM). The degree distribution, $\{P(k)\}_{k\in\mathbb{N}}$, is the distribution of the number of connections (the degree k) that nodes have. The degree-degree correlations are defined through the *joint degree distribution*, $\{P(k,k')\}_{k,k'\in\mathbb{N}}$, giving the probability that a randomly chosen edge has nodes of degree k and k' at its ends.

For both models, the size of the giant component S and the percolation threshold T_c can be calculated in the limit $N \to \infty$ using probability generating functions (pgf) [6, 10, 135, 136, 137, 147, 192, 193]. To model bond percolation on a given network with these models, we simply extract the degree distribution and the joint degree distribution; the required information therefore scales as k_max and k_max^2. The original network is then found within the random ensembles containing all possible networks that can be designed with the same degree distribution and/or degree-degree correlations. The readers unfamiliar with these models and/or the mathematics involved can get a brief overview of these subjects in Appendices A.7 and A.8.

The degree distribution and the joint degree distribution can be seen as the one-point and two-point correlation functions of a network. The next logical step would therefore be to consider three-point correlations (i.e., clustering), and eventually to incorporate mesoscopic features such as motifs, cliques, and communities. Although many theoretical models have

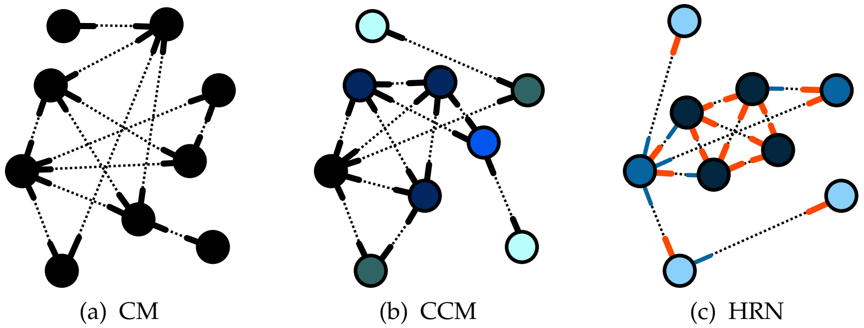


FIGURE A.1 – Comparison of the three random network models considered. (a) The CM randomly connects stubs drawn from a given degree distribution $\{P(k)\}_{k\in\mathbb{N}}$. (b) The CCM distinguishes nodes according to their degree (colors) and randomly match stubs according to the joint degree distribution $\{P(k,k')\}_{k,k'\in\mathbb{N}}$. (c) The HRN model distinguishes nodes by their coreness (colors) and stubs by their contribution to a node's coreness (thicker red or smaller blue stubs). Stubs are then randomly matched according to the matrices **K** and **C**.

been proposed [6, 23, 80, 98, 127, 138, 143, 176, 177, 180, 202], a general, objective, and systematic method to tune these models in order to reproduce the features found in real networks as well as to predict the outcome of bond percolation is yet to be found [3].

A.5 Hard-core Random Networks (HRN)

We propose an alternative approach by considering a macroscopic measure of centrality: the *coreness* of nodes. This choice is motivated by the recent observation that a node's coreness is a better indicator of the likeliness for that node to be part of the giant component than its degree [104]. This measure also has the advantage of being general, objective, systematic, and easily calculated [20].

A.5.1 Network coreness

The coreness c of a node is specified through its position in the core decomposition of a network [57, 174]. This decomposition assigns nodes to nested cores where nodes belonging to the n-th core all share at least n edges with one another. A node has a coreness equal to c if it is found in the c-th core, but not in the $(c+1)$-th core. The set of nodes with a coreness equal to c forms the c-shell.

This definition of the coreness may appear complicated to compute, but a simple algorithm allows us to do the decomposition very efficiently [20].

```
**Input** graph as lists of nodes $\mathcal{V}$ and neighbors $\mathcal{N}$
```

[3] Recent advances in understanding the global organization of clustering in real networks [46] offers further ideas to incorporate clustering in our model and will be the subject of a subsequent study.

```
Output list $\mathcal{C}$ with coreness for each node
compute and list the degrees $\mathcal{D}$ of nodes;
sort $\mathcal{V}$ with increasing degree of nodes;
for all $v \in \mathcal{V}$ in the order of $\mathcal{V}$ do
    $\mathcal{C}(v) := \mathcal{D}(v)$;
    for all $u \in \mathcal{N}(v)$ do
        if $\mathcal{D}(u) > \mathcal{D}(v)$ then
            $\mathcal{D}(u) := \mathcal{D}(u) - 1$;
        end if
    end for
    re-sort $\mathcal{V}$ accordingly
end for
```

In short, this algorithm is similar to a *pruning* process which removes nodes in order of their effective degree, i.e., their number of links shared with nodes currently ranked higher in the process. In the end, the coreness of a node is simply given by its degree once the peeling process reaches this particular node. Hence, we know that a node of degree k and coreness c has c *contributing* edges and $k - c$ *non-contributing* edges. Based on this key observation, we develop a coreness-based random network model that defines a maximally random network ensemble with an arbitrary degree distribution *and* an arbitrary core structure.

A.5.2 The HRN model

The only two inputs of the HRN model are a $\mathbf{K}$ matrix whose elements K_{ck} correspond to the fraction of the nodes that have a coreness c and a degree k, and a matrix $\mathbf{C}$ whose elements $C_{cc'}$ give the fraction of edges that leave nodes of coreness c to nodes of coreness c'. As this model considers undirected networks, the matrix $\mathbf{C}$ is symmetric and each edge is counted twice to account for both directions.

The HRN model is a multitype version of the CM [6, 10, 8] in which each node is assigned to a type, its coreness, and in which edges are formed by randomly pairing stubs that either contribute to the node's coreness (say, *red* stubs) or do not contribute to it (say, *blue* stubs). Red stubs from nodes of coreness c may be paired with blue stubs from nodes of coreness $c' \geq c$, or with red stubs attached to nodes of coreness $c' = c$ (intra-shell). Blue stubs stemming from nodes of coreness c may only be matched with red stubs stemming from nodes with a coreness $c' \leq c$. Blue stubs may never be paired together.

These rules enforce a minimal core structure, although random variations can bring nodes to a higher coreness than originally intended. For example, 3 nodes of original state $(k = 2, c = 1)$ could end up in the 2-shell in the unlikely event that they form a triangle. However, such random variations may never pull nodes to a lower coreness than intended, in addition to

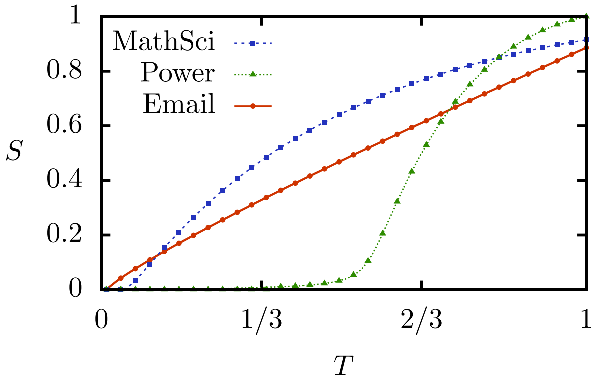


FIGURE A.2 – Validation of the HRN model. The predictions of Eqs.(A.9)–(A.10) (lines) are compared with the results obtained on networks generated with the Metropolis-Hastings algorithm described in Sec. A.5.3 (symbols). The matrices **K** and **C** were extracted from an email network, the MathSciNet co-authorship network and a power grid chosen for their different behaviors (see Table A.2 for datasets details). Numerical results (symbols) represent the average value of over $5 \cdot 10^5$ simulations performed on networks with more than $3 \cdot 10^5$ nodes.

being extremely unlikely in the limit of large networks ($N \to \infty$). The matrices **K** and **C** (see Appendix A.9 for consistency conditions) combined with the aforementioned stub pairing rules define a maximally random network ensemble with an arbitrary degree distribution and core structure (see Fig. A.1c).

The **K** matrix encodes several useful quantities. For instance, the fraction of nodes of coreness c

$$w_c = \sum_k K_{ck} \ , \tag{A.1}$$

and the associated joint degree distribution, i.e. the probability that a randomly chosen node of coreness c has k_r red stubs and k_b blue stubs

$$P_c(\boldsymbol{k}) \equiv P_c(k_r, k_b) = \frac{\delta_{c,k_r}}{w_c} K_{k_r, k_r + k_b} \ , \tag{A.2}$$

where δ_{c,k_r} is the Kronecker delta. Furthermore, we can extract the average degree of nodes of coreness c

$$\langle k \rangle_c = \frac{1}{w_c} \sum_k k K_{c,k} \tag{A.3}$$

and the average degree of the whole network

$$\langle k \rangle = \sum_{c,k} k K_{ck} \ . \tag{A.4}$$

It follows from the above definition that a fraction $w_c \langle k \rangle_c / \langle k \rangle$ of stubs stems from nodes of coreness c, of which a fraction $w_c c / \langle k \rangle$ is red and a fraction $w_c(\langle k \rangle_c - c)/\langle k \rangle$ is blue.

The **C** matrix encodes the transition probability $R(c', j|c, i)$ that a node of coreness c through a stub of color i [red (r) or blue (b)] leads to a node of coreness c' through one of its stubs of color j. Since inter-shell edges can only be formed by matching a red with a blue stub, we readily obtain

$$R(c', b|c, r) = \frac{C_{cc'}}{w_c c / \langle k \rangle} \tag{A.5a}$$

$$R(c, r|c', b) = \frac{C_{cc'}}{w_c(\langle k \rangle_c - c) / \langle k \rangle} \tag{A.5b}$$

$$R(c', r|c, b) = R(c, b|c', r) = 0 \tag{A.5c}$$

for $c < c'$. Similarly, as the pairing of blue stubs is forbidden [$R(c', b|c, b) = 0$ for any c and c'], a blue stub stemming from a node of coreness c leads to a node belonging to the same shell (through its red stub) with probability

$$R(c, r|c, b) = \frac{w_c(\langle k \rangle_c - c) / \langle k \rangle - \sum_{c''<c} C_{cc''}}{w_c(\langle k \rangle_c - c) / \langle k \rangle} \,. \tag{A.5d}$$

This last result is computed by subtracting the number of blue stubs leading to outer shells (i.e., lower coreness) to the total number of blue stubs stemming from nodes of coreness c, and then by normalizing [$\sum_{c',j} R(c', j|c, i) = 1$ for $c \in \mathbb{N}$ and $i \in \{r, b\}$]. Finally, symmetry with Eq. (A.5d) implies that

$$R(c, b|c, r) = \frac{w_c(\langle k \rangle_c - c) / \langle k \rangle - \sum_{c''<c} C_{cc''}}{w_c c / \langle k \rangle} \,, \tag{A.5e}$$

and normalization leads to

$$R(c, r|c, r) = \frac{2 w_c c / \langle k \rangle - C_{cc} - 2 \sum_{c''>c} C_{cc''}}{w_c c / \langle k \rangle} \,, \tag{A.5f}$$

where we have used the fact that $\sum_{c''} C_{cc''} = w_c \langle k \rangle_c / \langle k \rangle$.

To compute the size of the giant component in the limit of large networks ($N \to \infty$), we define a probability generating function (pgf)

$$g_c(\boldsymbol{x}) = \sum_{\boldsymbol{k}} P_c(\boldsymbol{k}) \prod_i \Big[(1 - T) + T \sum_{c',j} R(c', j|c, i) x_{c'j} \Big]^{k_i} \tag{A.6}$$

that generates the distribution of the number of nodes of each type (i.e., coreness c') that can be reached from a node of coreness c (the subscript j of the variable $x_{c'j}$ indicates the color of the stubs from which the node has been reached). To understand this last equation, consider a stub of color i stemming from a node of coreness c. This stub leads to an edge that has been removed with probability $1 - T$, or leads to a node of coreness c' through one of its stubs of color j with probability $TR(c', j|c, i)$. Since both the stub pairing and the edge removal are done randomly and independently, the distribution of the number of nodes that are neighbors of a node of coreness c having k_i stubs of color i is generated by the pgf

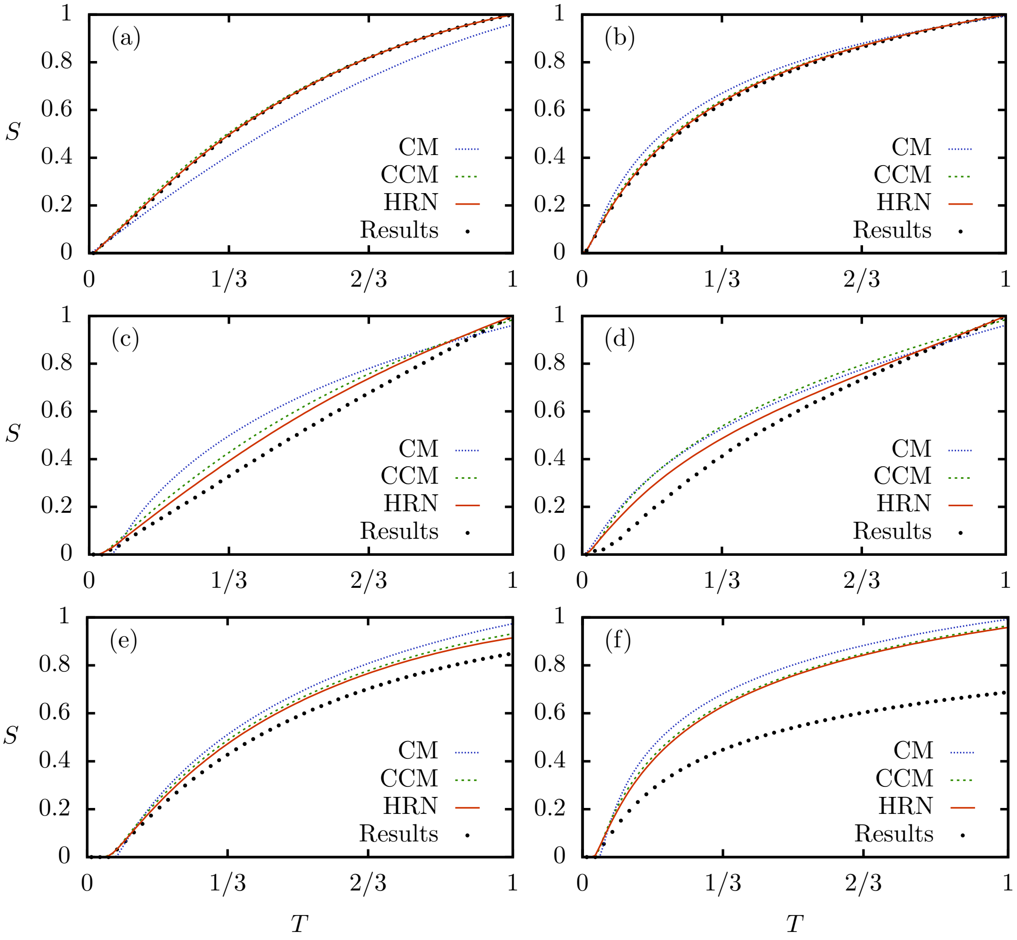


FIGURE A.3 – Results of bond percolation on real networks (black dots) compared with analytical predictions obtained with the CM (short dashed blue), CCM (long dashed green) and HRN (full red). The networks are: (a) Internet at the level of autonomous systems, (b) a snapshot of the Gowalla location-based social network, (c) the Pretty-Good-Privacy trust network, (d) a subset of the World-Wide Web, (e) the co-authorship network of MathSciNet before 2008, and (f) a large subset of the Facebook social network. See Table A.2 for further details.

$[(1-T)+T\sum_{c',j} R(c',j|c,i)x_{c'j}]^{k_i}$, a multinomial distribution. Multiplying the pgfs for both stub colors (neighborhood from stubs of different colors are also independent) and averaging over the distribution of the number of stubs of each color that nodes of coreness c have, $P_c(\boldsymbol{k})$, leads to Eq. (A.6).

Similarly, if the node had previously been reached via one of its red stubs, the distribution of its neighbors reachable via its *other* stubs—its *excess* degree distribution—is simply generated by the pgf

$$f_{cr}(\boldsymbol{x}) = \sum_{\boldsymbol{k}} P_c(\boldsymbol{k}) \prod_i \left[1 - T + T\sum_{c',j} R(c',j|c,i)x_{c'j}\right]^{k_i - \delta_{ir}} . \tag{A.7}$$

Finally, if the node had been reached via one of its blue stubs instead, the distribution of its neighbors reachable via its other stubs is generated by the pgf

$$f_{cb}(\boldsymbol{x}) = \sum_{\boldsymbol{k}} \frac{k_b P_c(\boldsymbol{k})}{\langle k \rangle_c - c} \prod_i \left[1 - T + T \sum_{c',j} R(c',j|c,i) x_{c'j}\right]^{k_i - \delta_{ib}} . \tag{A.8}$$

In this case, the probability over which $[(1-T) + T\sum_{c',j} R(c',j|c,i)x_{c'j}]^{k_i}$ must be averaged is weighted by the number of blue stubs that the node has (the denominator $\langle k \rangle_c - c$ is simply the normalization constant). For instance, since stubs are paired randomly, a randomly chosen blue stub is ten times more likely to belong to a node that has ten blue stubs than a node that has one.

These pgfs in hand, the size of the giant component is given by (see Ref. [8] for a complete and more general theoretical framework)

$$S = 1 - \sum_c w_c g_c(\boldsymbol{a}) \tag{A.9}$$

where $\boldsymbol{a} \equiv \{a_{ci}\}_{c\in\mathbb{N}, i\in\{r,b\}}$ is the probability that a node of coreness c reached by one of its stubs of color i does not belong to the giant component. More precisely, a node of coreness c belongs to the giant component if at least one of its neighbors belongs to it, which happens with probability $1 - g_c(\boldsymbol{a})$. The size of the giant component is then obtained by averaging this probability over the fraction of nodes that are of coreness c. The probabilities $\boldsymbol{a} \equiv \{a_{ci}\}_{c\in\mathbb{N}, i\in\{r,b\}}$ are obtained through a self-consistency argument: if a node of coreness c reached via one of its stubs of color i does not belong to the giant component, then neither should the nodes that can be reached from it. Hence these probabilties correspond to the stable fixed point of the system of equations

$$a_{ci} = f_{ci}(\boldsymbol{a}) \tag{A.10}$$

with $c \in \mathbb{N}$ and $i \in \{r, b\}$. As the distributions generated by $f_{ci}(\boldsymbol{x})$ are normalized, $\boldsymbol{a} = \boldsymbol{1}$ is always a solution of Eqs (A.10) and corresponds to the subcritical regime $S = 0$. At $T = T_\mathrm{c}$, this fixed point undergoes a transcritical bifurcation and looses its stability to another solution in $[0,1)^{c_{\max}}$. This supercritical regime corresponds to the existence of a giant component ($S > 0$); the critical point T_c is obtained from a stability analysis of Eqs. (A.10) around $\boldsymbol{a} = \boldsymbol{1}$.

A.5.3 Numerical HRN networks

To generate networks with a given core structure, we start with $N \gg 1$ nodes whose number of stubs is drawn from the degree distribution $\{P(k)\}_{k\in\mathbb{N}} = \{\sum_c K_{ck}\}_{k\in\mathbb{N}}$, and randomly match stubs to create edges (as done for the CM [136]). Next, for each node, we assign a coreness c with probability $Q_k(c) = K_{ck}/P(k)$; c of its stubs are then randomly selected as red and the $k - c$ others are identified as blue. Finally, we apply the following Metropolis-Hastings rewiring algorithm (similar to the one proposed in Ref. [135]). At each step, two

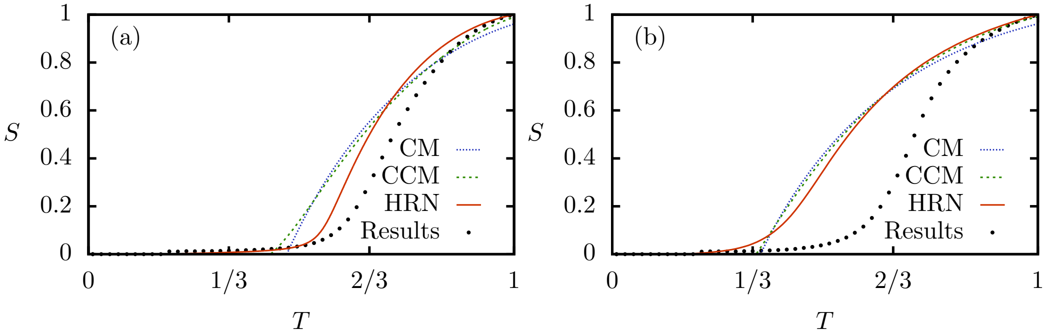


FIGURE A.4 – Results of bond percolation on real networks (black dots) compared with analytical predictions obtained with the CM (short dashed blue), CCM (long dashed green) and HRN (full red). The networks are: (a) a subset of the power grid of Poland, and (b) the Western States Power Grid of the United States. See Table A.2 for further details.

edges are randomly selected: edge 1 joins nodes of coreness c_1 and c_1' via their respective stubs of color i_1 and j_1 (c_2, i_2, c_2' and j_2 for edge 2). We replace these two edges by edge 3 (c_1, i_1, c_2 and i_2) and edge 4 (c_1', j_1, c_2' and j_2) with probability

$$\min\left\{1, \frac{\Gamma(c_1,i_1;c_2,i_2)\Gamma(c_1',j_1;c_2',j_2)}{\Gamma(c_1,i_1;c_1',j_1)\Gamma(c_2,i_2;c_2',j_2)}\right\} ,$$

where $\Gamma(c,i;c',j)$ is the wanted fraction of edges that join nodes of coreness c and c' via their respective stubs of color i and j. These fractions are readily obtained from the matrix $\mathbf{C}$ [joint probabilities of Eqs. (A.5)]

$$\begin{aligned} \Gamma(c,r;c',b) = \Gamma(c',b;c,r) &= C_{cc'} \\ \Gamma(c,r;c,b) = \Gamma(c,b;c,r) &= w_c(\langle k\rangle_c - c)/\langle k\rangle - \sum_{c''<c} C_{cc''} \\ \Gamma(c,r;c,r) &= 2w_c c/\langle k\rangle - C_{cc} - 2\sum_{c''>c} C_{cc''} \end{aligned} \tag{A.11}$$

where $c < c'$, and $\Gamma(c,i;c',j)$ is zero for all other combinations. This procedure preserves the degree distribution, and up to finite-size constraints, has the wanted core structure as its fixed point and is ergodic over the ensemble of networks defined by the HRN model. Figure A.2 compares the predictions of Eqs. (A.9)–(A.10) with the size of the giant component found in networks generated through this algorithm and shows a perfect agreement.

Description	N	$\langle k \rangle$	$k_{\max}$	$c_{\max}$	Fig.	Ref.
Web of trust of the Pretty Good Privacy (PGP) encryption algorithm	10 680	4.55	205	31	A.3(c), A.5	[26]
Structure of the Internet at the level of autonomous systems	22 963	4.22	2390	25	A.3(a),A.5	[88]
Large subset of the Facebook social network	63 891	5.74	223	16	A.3(f), A.5	[196]
Snapshot of the Gowalla location-based social network	196 591	9.67	14 730	51	A.3(b), A.5	[41]
Email exchange network from an undisclosed European institution	300 069	2.80	7 631	31	A.2, A.5	[109]
Subset of the World Wide Web	325 729	6.69	10 721	155	A.3(d), A.5	[15]
Co-authorship network of MathSciNet before 2008	391 529	4.46	496	24	A.2, A.3(e), A.5	[155]
Subset of the power grid of Poland	3 374	2.41	11	5	A.2, A.4(a), A.5	[201]
Western States Power Grid of the United States	4 941	2.67	19	5	A.4(b), A.5	[197]
Email communication within the University Rovira i Virgili	1 134	9.07	1 080	8	A.5	[153]
Protein-protein interactions in *S. cerevisiae*	2 640	5.00	111	8	A.5	[153]
Word association graph from the South Florida Free Association norms	7 207	8.82	218	7	A.5	[153]
Network of hyperlinks between Google's webpages	15 763	18.96	11 401	102	A.5	[155]
Reply network of the social news website Digg	30 398	5.60	283	9	A.5	[51]
The cond-mat arXiv co-authorship network circa 2005	30 561	8.24	191	15	A.5	[153]
Snapshot of the Gnutella peer-to-peer network	36 682	4.82	55	7	A.5	[166]
Email interchanges between different Enron email addresses	36 692	10.02	1 383	43	A.5	[106]
Brightkite location-based online social network	58 228	7.35	1 134	52	A.5	[41]
Network of tagged relationships on the Slashdot news website	77 360	12.13	2 539	54	A.5	[110]
Friendships between 100 000 Myspace accounts	100 000	16.82	59 108	78	A.5	[3]
Network of interactions between the users of the English Wikipedia	138 592	10.33	10 715	55	A.5	[115]
Co-acting network in movies released after December 31st 1999	716 463	21.40	4625	192	A.5	[88]

TABLE A.2 – Description and properties of the real networks used in Figs. A.2–A.5.

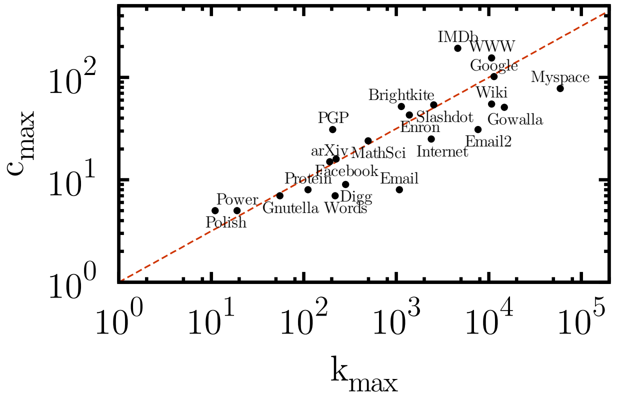


FIGURE A.5 – Relation between the highest coreness, $c_{\max}$, and the highest degree, $k_{\max}$, for different real networks. The dashed line corresponds to $c_{\max} \propto \sqrt{k_{\max}}$.

A.5.4 Results

Figures A.3–A.4 display the predictions of Eqs. (A.9)–(A.10) with the size of the giant component found in real networks (see caption and Table A.2 for a complete description), and with the predictions of the CM and the CCM. These particular networks were chosen to highlight some important results.

We find that the HRN model performs *at least as well* as the CCM in all investigated cases. First, this observation is interesting as the HRN model requires less input information than the CCM. Indeed the required information scales roughly as $k_{\max}c_{\max} + c_{\max}^2$. As shown in Fig. A.5, $c_{\max}$ scales approximately as $k_{\max}^{1/2}$ in many real networks, hence the input information in the HRN model scales roughly as $k_{\max}^{3/2}$. Considering the fact that $k_{\max}$ in real networks is often well above 10^2 (see Table A.2), this difference results in a much faster computation and a major memory gain.

Second, although the HRN model does not account explicitly for the degree-degree correlations, they are effectively captured by the matrices **K** (degree-coreness correlations) and **C** (coreness-coreness correlations). This is confirmed in all available real networks as seen in Figs. A.3–A.4 and in Table A.3 (see the correlation coefficient r). Table A.3 further investigates the efficiency of the HRN model to reproduce the structural properties of real networks by comparing the clustering coefficients and the mean shortest paths found in both the real networks and their equivalent HRN networks. As expected for any random networks, the HRN model has vanishing clustering. Its relatively good performance for the Internet is a mere consequence of the small size of some of the c-shells. Furthermore, the HRN model once more performs at least as well as the CCM in reproducing the mean shortest path. Interestingly, we have found empirically that the closer ℓ is to the value obtained in the real

Network	C	r	ℓ/ℓ^{CCM}
Internet	0.230	-0.099	1.04
Internet HRN	0.114	-0.081	1.00
PGP	0.266	0.490	1.39
PGP HRN	0.011	0.511	1.14
Polish Grid	0.019	0.677	1.55
Polish Grid HRN	0.002	0.527	1.25

TABLE A.3 – Comparison of the clustering coefficient C [197], the degree correlation coefficient r [135] and the mean shortest path ℓ [197] found in three real networks considered in this paper and in the equivalent HRN networks of the same size generated using the algorithm presented in Sec. A.5.3. The three networks correspond to a representative sample of the different behaviors observed with the real networks presented in Table A.2. The mean shortest paths are expressed as a ratio with the mean shortest paths obtained in the equivalent CCM networks ($\ell^{\mathrm{CCM}} = 3.68$, 5.40 and 9.44 for the Internet, PGP and Polish power grid networks respectively).

network, the better the HRN model predicts percolation.

Moreover, and perhaps surprisingly, we see in Fig. A.4(a) that the "S" shape obtained from the Polish power grid, typically due to finite size, is well reproduced by the HRN model, which is formally infinite in size. More precisely, this shape is usually attributed to the finite size of the network ($N = 3374$ for the Polish power grid) as the small components—whose average size formally diverges at $T = T_c$—are misinterpreted as an emerging giant component. Interestingly, the results from the HRN model suggest that this shape is not a numerical artifact of the percolation algorithm, but that it is rather a signature of its geographically-embedded nature due to strong *coreness-related* correlations. This unexpected property of the HRN model is confirmed on another, more clustered, Western States power grid on Fig. A.4(b). In this case, adding clustering to the HRN is expected to shift its prediction towards higher values of T, i.e., closer to the results from the real network. In fact, the HRN model is more accurate in predicting percolation on the Polish power grid (clustering coefficient $C = 0.02$) than for the Western States power grid ($C = 0.08$). A clustered version of the HRN model seems to offer a promising avenue for the modeling of geographically-embedded networks such as power grids.

In this regard, the results of Figs. A.3(e)–(f) add even more emphasis on the importance of including the effect of clustering in a subsequent version of the HRN model. Indeed, co-authorship networks A.3(e) are notoriously clustered networks as authors of a same paper are all connected via a fully-connected clique. Similarly, in Facebook A.3(f), people belonging to a same social group (e.g., classmates, colleagues, teammates) tend all to be connected to one another, yielding almost fully-connected cliques. Again, we expect in this situation that clustering would reduce the size of the giant component (due to redundant connections in cliques), hence bringing the predictions of a clustered HRN model closer to the behaviors observed with the real networks.

A.6 Conclusion

We have shown that the core structure can be useful beyond the characterization and visualization of networks. It serves well modeling efforts and is efficient in reproducing the structural properties of real networks. Moreover, a few simple connection rules can enforce a core structure in random networks for which the outcome of bond percolation can be predicted with the well-established pgf approach. We feel that this work sets the stage for further improvements (specifically the inclusion of clustering) and paves the way towards a more complete analytical description of percolation on real networks.

A.7 Appendix: Configuration Model

The most influential quantity with regard to bond percolation on networks is the degree distribution: the distribution of the number of connections (degree) that nodes have. The simplest analytical model that incorporates an arbitrary degree distribution is the CM [136, 147]. It defines a maximally random network ensemble that is random in all respects other than the degree distribution $\{P(k)\}_{k\in\mathbb{N}}$: the probability for a randomly chosen node to have a degree equal to k. Networks of this ensemble are generated by creating a set of N nodes, each with a number of stubs drawn from the degree distribution, and then by pairing randomly stubs to form edges.

To compute the size S^{CM} of the giant component and the value T_c^{CM} of the percolation threshold, we define the probability generating function [136]

$$g(x) = \sum_{k=0}^{\infty} P(k)[(1-T)+Tx]^k \tag{A.12}$$

that generates the degree distribution. The first derivative of $g(x)$ evaluated at $x = 1$ corresponds to the average degree of the nodes $g'(1) = \langle k \rangle$. We also define

$$f(x) = \frac{g'(x)}{g'(1)} = \frac{1}{\langle k \rangle} \sum_{k'=1}^{\infty} k' P(k')[(1-T)+Tx]^{k'-1} \tag{A.13}$$

that generates the number of *other* neighbors of a node that has been reached by following a randomly chosen edge (i.e., the *excess* degree distribution). The size of the giant component is directly obtained via

$$S^{\mathrm{CM}} = 1 - g(a^{\mathrm{CM}}) \,, \tag{A.14}$$

where a^{CM} is the probability that a randomly chosen edge does not lead to the giant component. It is the stable fixed point of

$$a^{\mathrm{CM}} = f(a^{\mathrm{CM}}) \tag{A.15}$$

in $[0,1]$. The solution $a^{\mathrm{CM}} = 1$ corresponds to the absence of a giant component ($S^{\mathrm{CM}} = 0$). The percolation threshold is the point at which this solution becomes unstable.

To model bond percolation on a given network with the CM, one simply has to extract the degree distribution; the required information therefore scales as k_{max}, the highest degree of the network. The original network is then found within the network ensemble generated by the CM, the ensemble composed of all possible networks one could design with the exact same degree distribution.

A.8 Appendix: Correlated Configuration Model

Apart from the degree distribution, real networks are typically characterized by strong correlations regarding *who is connected with whom*. One way to include such correlations into a random network model is through the *joint degree distribution* $\{P(k,k')\}_{k,k'\in\mathbb{N}}$ giving the probability that a randomly chosen edge has nodes of degree k and k' at its ends. This yields a *Correlated Configuration Model* (CCM) that defines a maximally random network ensemble having arbitrary degree-degree correlations with a corresponding degree distribution [135, 193]. The degree distribution is encoded in $\{P(k,k')\}_{k,k'\in\mathbb{N}}$ through the identity

$$\sum_{k'} P(k,k') = \frac{kP(k)}{\langle k \rangle} \,. \tag{A.16}$$

Generating networks from this ensemble proceeds as for the CM: N nodes, whose degrees are drawn from $\{P(k)\}_{k\in\mathbb{N}}$, are connected via the stub pairing scheme. A Metropolis-Hastings rewiring algorithm [135] is then applied whose fixed point is the network ensemble defined by $\{P(k,k')\}_{k,k'\in\mathbb{N}}$. At each step, two edges are randomly chosen: edge 1 joins nodes m_1 and n_1 with respective degree i_1 and j_1 (m_2, n_2, i_2 and j_2 for edge 2). These two edges are replaced by edge 3 (m_1, m_2, i_1 and i_2) and edge 4 (n_1, n_2, j_1 and j_2) with probability

$$\min\left\{1, \frac{P(i_1,i_2)P(j_1,j_2)}{P(i_1,j_1)P(i_2,j_2)}\right\} \,. \tag{A.17}$$

The size S^{CCM} of the giant component is computed as in the CM [135]

$$S^{\mathrm{CCM}} = 1 - \sum_{k=0}^{\infty} P(k)[(1-T) + Ta_k]^k = 1 - g(\boldsymbol{a}) \tag{A.18}$$

where $\boldsymbol{a} = \{a_k\}_{k\in\mathbb{N}}$ is the set of probabilities that an edge leading toward a node with a degree k is not attached to the giant component. They correspond to the stable fixed point in

$[0,1]^{k_{\max}}$ of the system of equations

$$a_k = \frac{\sum_{k'} P(k,k')[(1-T)+Ta_{k'}]^{k'-1}}{\sum_{k'} P(k,k')} \tag{A.19}$$

with $k \in \mathbb{N}$. The value T_c of the percolation threshold is the value for which the fixed point $\boldsymbol{a} = \mathbf{1}$ of Eqs. (A.19) becomes unstable.

To model bond percolation on a given network with the CCM, one simply has to extract the joint degree distribution. This is achieved by scanning the degree of the two nodes at the end of each edge of the network; the required information therefore scales as $k^2_{\max}$. The original network is then found within the random network ensemble of all networks with the same degree distribution and degree-degree correlations. Note that this ensemble is a subset of the ensemble generated by the CM with the same degree distribution.

A.9 Appendix: Consistency conditions on K and C

The consistency conditions on the matrices **K** and **C** can be summarized as follows: they must encode an ensemble of *closed* networks. In other words, *all stubs must be paired*, and this must be done in accordance with the stubs matching rules (e.g., two blue stubs cannot be paired). Consequently, there is no k-core structure that the HRN model cannot model as long as it is realistic. This will always be the case when **K** and **C** are extracted from real networks.

First, there must be as many edges leaving nodes of coreness c toward nodes of coreness c' as there are in the opposite direction. This requires that $C_{cc'} = C_{c'c}$, a condition that is always fulfilled since **C** is defined as a symmetric matrix

$$\mathbf{C} = \mathbf{C}^{\mathrm{T}} \,. \tag{A.20}$$

Secondly, the degree of each node is bounded from below by its coreness, hence

$$K_{ck} = 0 \qquad \text{for } k < c \,. \tag{A.21}$$

Thirdly, both **K** and **C** must prescribe the same number of stubs stemming from nodes of coreness c

$$\sum_k kK_{ck} = \langle k \rangle \sum_{c'} C_{cc'} \,, \tag{A.22}$$

where the extra factor $\langle k \rangle$ accounts for the fact that **K** "counts" nodes whereas **C** "counts" stubs (i.e., multiplying both sides by the number of nodes N yields absolute numbers instead of *per capita* averages). Finally, as the coreness of the nodes defines their number of red stubs, the matrix **C** is subjected to the following additional constraints for every c

$$\langle k \rangle \sum_{c'>c} C_{cc'} \leq w_c c \leq \langle k \rangle \sum_{c' \geq c} C_{cc'} \, . \tag{A.23}$$

The first inequality states that there must be at least as many red stubs stemming from nodes of coreness c as there are edges leaving the c-shell toward nodes of higher coreness. Equality then means that all red stubs lead to node of higher coreness. The second inequality states that all red stubs must lead to nodes of coreness c or higher. Equality occurs when all blue stubs are directed toward nodes of coreness $c' < c$. A similar expression to (A.23) can be derived for blue stubs

$$\langle k \rangle \sum_{c'<c} C_{cc'} \leq (\langle k \rangle_c - c) w_c \leq \langle k \rangle \sum_{c' \leq c} C_{cc'} \, , \tag{A.24}$$

and can be interpreted analogously.

Annexe B

Application III : Stratégies locales d'immunisation sur graphes

Article original :

Global efficiency of local immunization on complex networks

Laurent Hébert-Dufresne [1], **Antoine Allard**[1], Jean-Gabriel Young[1] et Louis J. Dubé

Département de Physique, de Génie Physique, et d'Optique,
Université Laval, Québec (Qc), Canada G1V 0A6

Référence : Sci. Rep. **3**, 2171 (2013) [91]

Distribué selon : Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License (§B.2 –§B.9) [2]

[1] Le développement théorique, les simulations numériques de même que la rédaction du manuscrit sont essentiellement les fruits du travail des trois premiers auteurs. Le dernier auteur a participé à l'élaboration générale du projet de même qu'à la révision du manuscrit.

[2] Ces sections contiennent le contenu original de l'article. Celui-ci n'a été modifié que pour se conformer au format exigé par la Faculté des études supérieures et postdoctorales de l'Université Laval.

B.1 Avant-propos

Nous terminons cette thèse en présentant une troisième application du modèle décrit au chapitre 4. Ce projet prend racine dans la thèse que Laurent Hébert-Dufresne avait mise de l'avant lors de ses travaux de maîtrise, c.-à-d. que la structure communautaire serait le niveau organisationnel fondamental des systèmes complexes [87]. Nous avions alors proposé en conclusion d'un article produit à l'époque que si cette thèse s'avérerait correcte, il existerait des nœuds appartenant à plusieurs communautés, des *structural hubs*, qui seraient des cibles de choix pour toute stratégie d'immunisation contre un agent infectieux [88].

À l'instar du projet présenté à l'annexe A, ce projet fut au départ de nature purement numérique : nous testions différentes stratégies d'immunisation sur des graphes de diverses natures à l'aide de simulations numériques intensives. Ces calculs nous permirent de tirer quelques conclusions intéressantes, dont une insoupçonnée qui mit en lumière un raccourci trompeur dans la nomenclature d'une étude qui avait fait grand bruit à l'époque [104]. En effet, les auteurs de cette étude avaient identifié les nœuds au cœur de la décomposition en couches (*highest coreness*) comme étant les propagateurs les plus influents (*influential spreaders*). Les auteurs s'appuyaient notamment sur le fait qu'à degré égal, les nœuds plus au cœur du graphe avaient une plus grande probabilité d'être ultimement infectés. Or, en définissant l'influence d'un nœud par l'impact que son immunisation a sur la taille d'une épidémie, nous avons d'une part montré que l'immunisation de ces nœuds dits influents avait peu d'impact sur la taille d'une épidémie, et que les auteurs avaient plutôt identifiés les nœuds qui étaient les plus *susceptibles* d'être infectés lors d'une éventuelle épidémie.

Cette étude nous a également permis de valider notre hypothèse de départ et d'en définir les conditions de validité. Les *structural hubs* étaient les nœuds les plus influents en ce qui a trait à l'issue d'une épidémie, mais seulement lorsque certaines conditions étaient réunies. De par leur nature, les simulations numériques rendaient difficile l'identification de ces conditions, et c'est à cet égard que l'approche théorique présentée dans cette thèse nous fut utile. En effet, en tentant de reproduire le comportement que les simulations numériques suggéraient (voir la figure B.8), nous avons été en mesure d'identifier correctement les paramètres pour lesquels la stratégie d'immunisation que nous mettions de l'avant était la plus efficace. Ainsi, bien qu'il joua un rôle plus modeste que dans le projet présenté au chapitre A, cet article rend compte de l'utilité du *laboratoire théorique* que nous avons développé, même lors d'études où la modélisation mathématique n'est pas l'objectif principal.

Quantity	Description	Definition
λ	Ratio of the infection rate α and the recovery rate β in the SIS dynamics	Sec. B.4.1
I^*	Prevalence of the disease in the SIS dynamics	Sec. B.4.1
T	Transmission probability in the SIR dynamics	Sec. B.4.1
R_f	Number of nodes that have been infected during a SIR epidemics	Sec. B.4.1
ε	Fraction of nodes that are fully immunized	Sec. B.4.1
k	Degree of a node	Sec. B.4.1
m	Number of communities to which a node belongs	Sec. B.4.1
b	Betweenness centrality of a node	Sec. B.4.1
c	Coreness of a node	Sec. B.4.1
ρ	Average density of the communities	Sec. B.4.2
M	Number of nodes in motifs	Sec. B.4.4
$p(i,j)$	Probability that a node belongs to i motifs and has j single links	Sec. B.4.4

TABLE B.1 – Glossary of the major mathematical objects defined in this chapter.

B.2 Abstract

Epidemics occur in all shapes and forms: infections propagating in our sparse sexual networks, rumours and diseases spreading through our much denser social interactions, or viruses circulating on the Internet. With the advent of large databases and efficient analysis algorithms, these processes can be better predicted and controlled. In this study, we use different characteristics of network organization to identify the influential spreaders in 17 empirical networks of diverse nature using 2 epidemic models. We find that a judicious choice of local measures, based either on the network's connectivity at a microscopic scale or on its community structure at a mesoscopic scale, compares favorably to global measures, such as betweenness centrality, in terms of efficiency, practicality and robustness. We also develop an analytical framework that highlights a transition in the characteristic scale of different epidemic regimes. This allows to decide which local measure should govern immunization in a given scenario.

B.3 Introduction

Epidemics never occur randomly. Instead, they follow the structured pathways formed by the interactions and connections of the host population [35, 100]. The spreading processes relevant to our everyday life take place on networks of all sorts: social (e.g. epidemics [12, 99]), technological (e.g. computer viruses [82, 157]) or ecological (cascading extinctions in food webs [60]). With a network representation, these completely different processes can be modelled as the propagation of a given agent on a set of nodes (the population) and links (the interactions). Different systems imply networks with different organizations, just as different agents require different epidemic models.

There has long been significant interest in identifying the *influential spreaders* in networks. Which nodes should be the target of immunization efforts in order to optimally protect the network against epidemics? Unfortunately, most studies feature two significant shortcomings. Firstly, the proposed methods are often based on optimization or heuristic algorithms requiring nearly perfect information on a static system (e.g. [39, 75]); this is rarely the case. Secondly, methods are usually tested on small numbers of real systems using a particular epidemic scenario (e.g. [120, 169]); this limits the scope of possible outcomes.

We first present a numerical study, perhaps the largest of its kind to date, where we argue that, depending on the nature of the network and of the disease, different immunization tactics have to be taken into consideration. In so doing, we formalize the notion of node influence and illustrate how *local knowledge* around a particular node is usually sufficient to estimate its role in an epidemic. We also show how, in certain cases, the influence of a node is not necessarily dictated by its number of connections, but rather by its role in the network's community structure (see Fig. B.1). Far from trivial, it follows that an efficient immunization

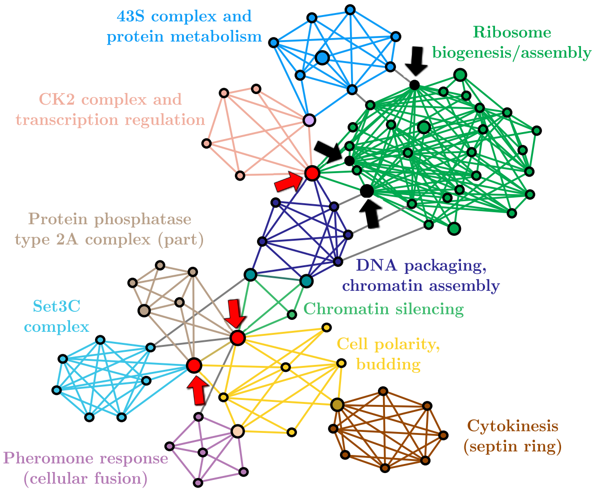

43S complex and protein metabolism
Ribosome biogenesis/assembly
CK2 complex and transcription regulation
Protein phosphatase type 2A complex (part)
DNA packaging, chromatin assembly
Chromatin silencing
Set3C complex
Cell polarity, budding
Cytokinesis (septin ring)
Pheromone response (cellular fusion)

FIGURE B.1 – Protein interactions of *S. cerevisiae* (subset) [153]. The three black nodes correspond to the ones with the highest degree, and the three red ones have the highest membership number. In this particular example, it is readily seen that the latter are structurally more influent.

strategy can be obtained solely from local measures, which are easily estimated in practice and robust to noisy or incomplete information. We further develop an analytical formalism ideally suited to test the effects of local immunization on realistic network structures. Combining the insights gathered from the numerical study and this formalism, we finally formulate a readily applicable approach which can easily be implemented in practice.

B.4 Results

B.4.1 Models and measures

There exist two standard models emulating diverse types of epidemics: the *susceptible-infectious-recovered* (SIR) and *susceptible-infectious-susceptible* (SIS) dynamics. In both, an infectious node has a given probability of eventually infecting each of its susceptible neighbors during its infectious period, which is terminated by either death/immunity leading to the recovered state (SIR) or by returning to a susceptible state (SIS). In the SIR dynamics, for a given transmission probability T, the quantity of interest is the mean fraction R_{f} of recovered nodes once a disease, not subject to a stochastic extinction, has finished spreading (i.e. we focus on the giant component [147]). Since each edge can only be followed once, this dynamics investigates how a population is vulnerable to the *invasion* of a new pathogen. In the SIS dynamics, we are interested in the prevalence I^* (fraction of infectious nodes) of the disease at equilibrium (equal amounts of infections and recoveries) as a function of the ratio $\lambda = \alpha/\beta$ of infection rate α and recovery rate β. This particular dynamics permits the study of how a given network structure can *sustain* an already established epidemic.

Should a fraction ε of the population be fully immunized, our objective is to identify the nodes whose absence would minimize R_f and I^*. The *epidemic influence* of a node — that is the effect of its removal on R_f and I^* — depends mainly on its role in the organization of the network. Hence to efficiently immunize a population, we must first understand its underlying structure.

Network organization can be characterized on different scales, each of which affect the dynamics of propagation. At the microscopic level, the most significant feature is the *degree* of a node (its number of links, noted k) which in turn defines the degree distribution of the network. The significance of the high-degree nodes (the *hubs*) for network structure in general [15], for network robustness to random failure [4] and for epidemic control [158] has long been recognized.

At the macroscopic level, the role of a node can be described by its *centrality*, which may be defined in various ways. Frequently used in the social sciences is the *betweenness centrality* (b), quantifying the contributions of a given node to the shortest paths between every pair of nodes in the network [71]. Arguably, this method should be among the best estimate of a node's epidemic influence as it directly measures its role in the different pathways between all other individuals [16], yet at a considerable computational cost. A simpler method, the k-core (or k-shell) decomposition [19, 20], assigns nodes to different layers (or *coreness* c) effectively defining the core and periphery of a network (high and low c respectively). It has recently been shown that coreness is well suited to identify nodes that are the most at risk of being infected during the course of an epidemic [104]. In light of our results, we will be able to discuss the distinction between a node's vulnerability to infection and its influence on the outcome of an epidemic.

The mesoscopic scale has recently been the subject of considerable attention. At this level of organization, the focus is on the redundancy of connections forming dense clusters referred to as the community structure of the network [2, 153]. Nodes can be distinguished by their *membership* number m, i.e., the number of communities to which they belong. We will consider that two links of a given node are part of one community if the neighbours they reach lead to significantly overlapping neighbourhoods [2]. This definition is directly relevant to epidemic dynamics as links within communities do *not* lead to new potential infections. We call *structural hubs* the nodes connecting the largest number of different communities. These nodes act as bridges facilitating the propagation of the disease from one dense cluster to another. Targeting structural hubs to hinder propagation in structured populations has been previously proposed and investigated [120, 169], but has yet to be tested extensively.

Note that the microscopic and mesoscopic levels (as defined above) are characterized by *local measures* in the sense that they do not require a complete knowledge of the network, in contrast to *global measures* like the betweenness centrality. Moreover, as we will see, *local* *measures are less sensitive to incomplete or incorrect information*. Adding, removing or rewiring a link only affects the degree or membership of nodes directly in the neighbourhood of the modification; whereas the same alterations can potentially affect the centrality of nodes anywhere in the network through cascading effects. Furthermore, even if community detection often requires the tuning of a global resolution parameter, we will see that this additional step does not affect the identification of structural hubs, meaning that local information is sufficient to accurately determine a node's memberships.

In our numerical simulations we will have a perfect knowledge of static networks. This will allow us to use global measures as a reference to test the efficiency of local measures best suited in practice. We therefore ask without discrimination: which of the degree k, the coreness c, the betweenness centrality b or the membership number m is the best identifier of the most influential nodes on the outcome of an epidemic? To answer this question, we have simulated SIR and SIS dynamics with Monte Carlo calculations on 17 real-world networks. In each case, a fraction ε of the nodes was removed in decreasing order of the nodes' score for each of the four different measures. By comparing their efficiency to reduce R_f or I^* as a function of ε, we are able to establish which measure is best suited for a given scenario characterized by a network structure, a propagation dynamics and a disease transmissibility (i.e. probability of transmission).

B.4.2 Case study: a data exchange network

We first illustrate our methods using the network of users of the Pretty-Good-Privacy algorithm for secure information interchange (hereafter, the PGP network) [26], which could be the host of the propagation of computer viruses, rumors or viral marketing campaigns. Results for the 16 other networks are presented and discussed in the next section as well as in the Supporting Information (SI) document.

Communities in the network are extracted with the link community algorithm of Ahn *et al.* [2]. This algorithm groups links — and therefore the nodes they join — into communities based on the overlap of their respective neighbouring nodes. It is this overlap that reduces the number of new potential infections in a community structure, as opposed to a random network. This method thus reflects our understanding of how communities affect disease propagation. While it may not directly detect the social groups or functional modules of a network, it identifies significant clusters of redundant links. This redundancy or overlap is quantified through a Jaccard coefficient, and two links are grouped into the same community when their coefficient exceeds a certain threshold. The threshold value acts as a resolution, enabling to look at different levels of organization. As suggested in [2], the value of the threshold is chosen to maximize the average density ρ of the communities (see Material and Methods). As this choice may seem arbitrary, Fig. B.2 investigates the similarity between the nodes with the highest membership numbers, for different thresholds. It suggests that the

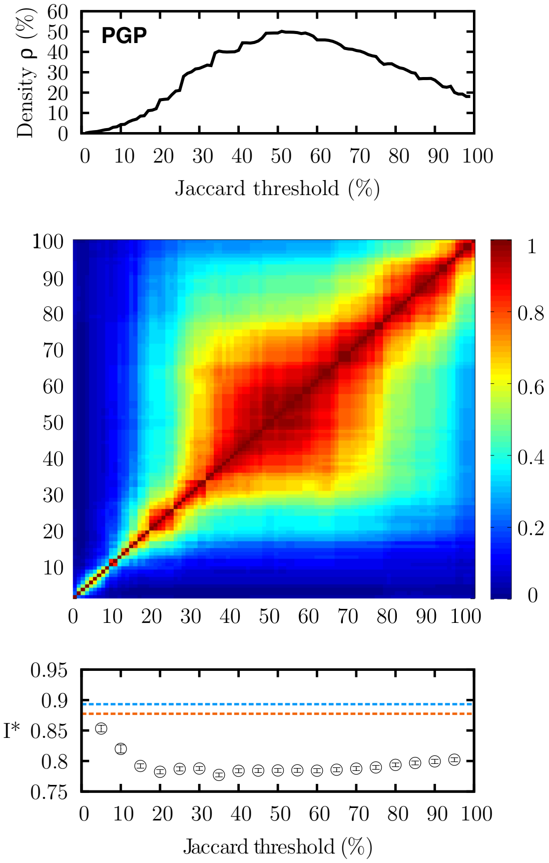


FIGURE B.2 – Robustness of structural hubs in the PGP network. (top) Community density (ρ) obtained through different Jaccard thresholds. (middle) Robustness of the structural hubs identification methods. Element (i, j) gives the overlap (normalized) between the structural hubs (top 1%) selected with thresholds i and j. The highest line and last column of the matrix correspond to the case where the membership number equals the degree. (bottom) Prevalence I^* of SIS epidemics with $\lambda = 5$ when the top 1% of structural hubs are removed (compared with the results without removal in blue or with random targets in orange).

membership number is fairly robust around the threshold. Moreover, Fig. B.2 also demonstrates that the effect of the removal of the structural hubs on a SIS epidemics is very robust to the choice of the threshold. Thus, we will henceforth use the membership numbers obtained with the threshold value corresponding to the highest community density.

The differences, if any, between the efficiency of the different methods are due to the immunized nodes not being the same. Figure B.3(top) investigates the correlations between the different properties (k, b, c and m) of each node. Perhaps the most important result here is that nodes with a high membership number may have relatively small degree, coreness and betweenness centrality. Hence, we expect the immunizing method based on community structure to have a different influence on the outcome of epidemics. Figure B.3(bottom) shows the consistensy (or lack thereof) of a given measure, depending on the quality of the

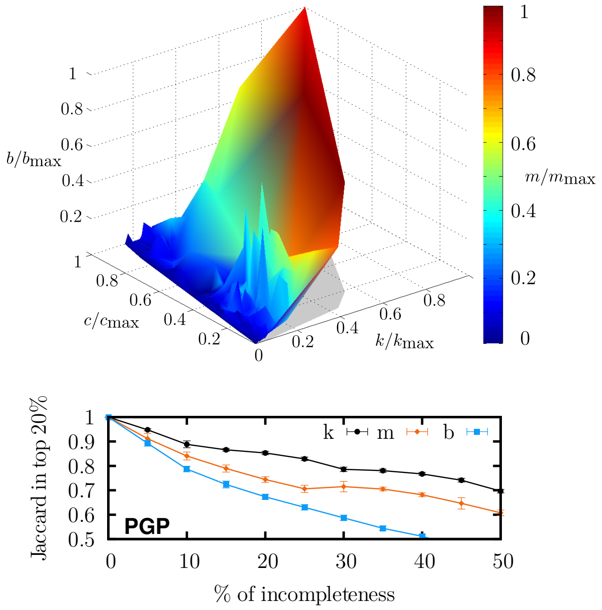


FIGURE B.3 – Difference in immunization targets for the PGP network. (top) We present correlations between the degree (k, right axis), the coreness (c, left axis), the betweenness centrality (b, vertical axis) and the membership number (m, color) for each nodes. Each measure is normalized according to the highest value found in the network. Each node is represented in this 4-dimensional space and a simple triangulation procedure then yields a more intelligible appearance. Structural hubs (dark red) can be found even at relatively small degree ($\sim k_{\max}/2$), coreness ($\sim c_{\max}/5$) and centrality ($\sim b_{\max}/3$). (bottom) Jaccard coefficient between the ensemble of nodes identified as part of the top 20% according to a given measure (k, m or b) on two versions of the network: the original complete network and a network ensemble where a certain percentage of links has been randomly removed (horizontal axis). The shorter the range of a measure, the more robust it is to incomplete information.

available data. The robustness of local (micro and meso) measures is of obvious practical advantage. Both robustness and correlations are further investigated in the SI.

To study various epidemic scenarios, we consider both SIS and SIR dynamics (which may behave quite differently) with different values of the transmission probability (λ and T for SIS and SIR, respectively). In fact, each network features an *epidemic threshold*, i.e. critical values λ_c [93] and T_c [136], below which I^* and R_f vanish to zero in an equivalent infinite network ensemble. As we will show, the observed behavior can differ significantly depending whether or not λ and T are close to their critical value.

Figure B.4 presents results of different immunization methods against SIS dynamics for different values of λ. On the top figure, where λ is near λ_c, the most successful method of intervention is to target nodes according to their degree. At low transmissibility, the dis-

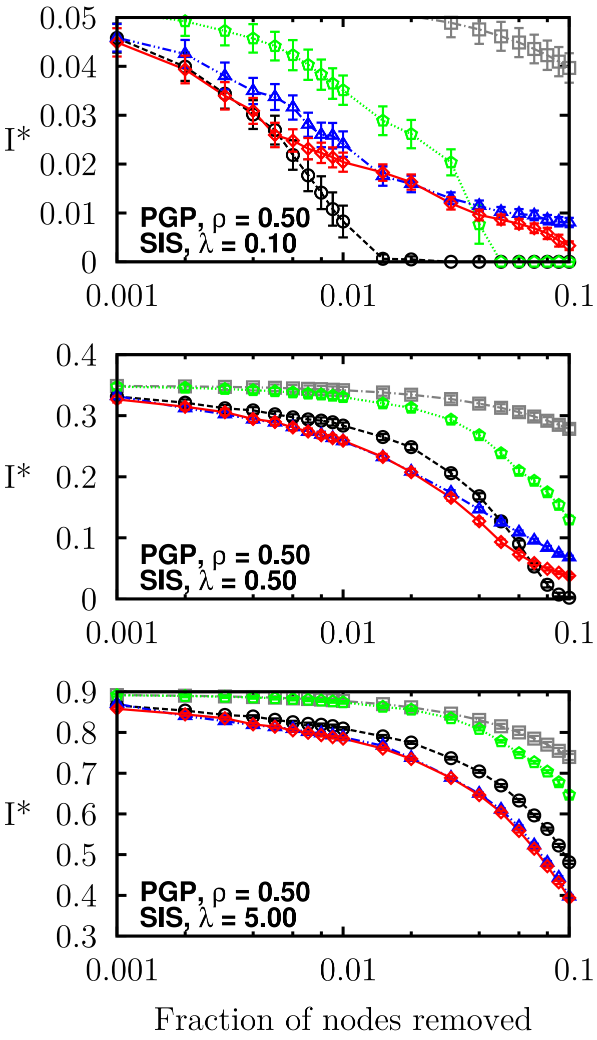


FIGURE B.4 – Efficiency of the immunization methods against an SIS epidemics on the PGP network. Nodes are removed in decreasing order of their score according to each method: coreness (green pentagons), degree (black circles), betweenness centrality (blue triangles) and memberships (red diamonds) and the effect of removal is then quantified in terms of the decrease of the prevalence I^*. The prevalence of the epidemics when the removed nodes are chosen at random (grey squares) has been added for comparison. Figures are presented in increasing order of transmissibility (λ) from top to bottom.

ease follows only a very small fraction of all links. The shortest paths are seldom used and the poor performance of betweenness centrality follows. Moreover, the disease will not be affected by the community structure, because even in dense neighbourhoods, most links will not be travelled. We then say that the disease, unaffected by link clustering, follows a tree-like structure (without loops), where community memberships are insignificant. It is therefore better to simply remove as many links as possible.

As λ increases beyond λ_c, we see that immunization based on membership numbers quickly

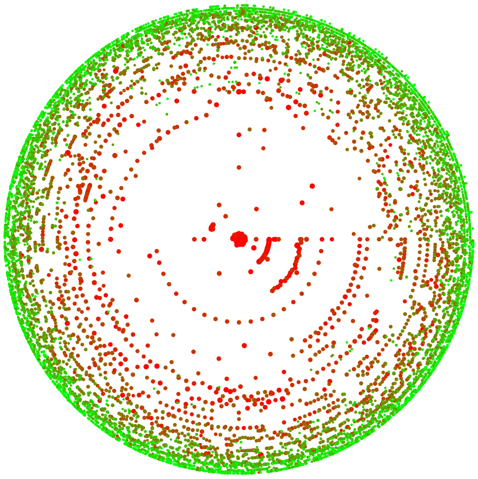


FIGURE B.5 – k-core decomposition of the PGP network. Representation (based on [11]) of the k-shells in the PGP network with nodes colored according to their total infectious period during a given time interval. Red nodes are more likely to be infectious at any given time than green nodes as the color is given by the square of the fraction of time spent in infectious state. Note how the central nodes (the core) of the network are most at risk.

outperforms the other methods. As more links are travelled, the disease is more likely to follow superfluous links in already infected communities. Hubs sharing their many links within few communities are therefore not as efficient in causing secondary infections as one might expect. Similarly, targeting through betweenness centrality also performs better with higher λ, albeit not as well as membership-targeting in this case. For $\lambda \gg \lambda_c$, immunization based on membership numbers (local) and on betweenness centrality (global) converge toward similar efficiency, significantly outperforming degree-based immunization.

Another interesting feature of our results is the poor performance of immunization based on node coreness. A previous study had clearly shown that epidemics mostly flourished within the core of the network (see Fig. B.5) because of its density [104]. Ironically, this density also implies redundancy. While the core nodes are highly at risk of being infected, their removal has a limited effect because there exist alternative paths within their neighbourhood: the core offers a perfect environment to the disease and is consequently robust to node removal. It is therefore more effective to stop the disease from reaching, or leaving, the core by removing the nodes bridging other neighbourhoods (i.e. the structural hubs).

Similar conclusions are drawn for the SIR dynamics. As T moves away from T_c, the most significant level of organisation shifts from the degree (microscopic) to communities (mesoscopic) as membership-based immunization progressively outperforms the other strategies.

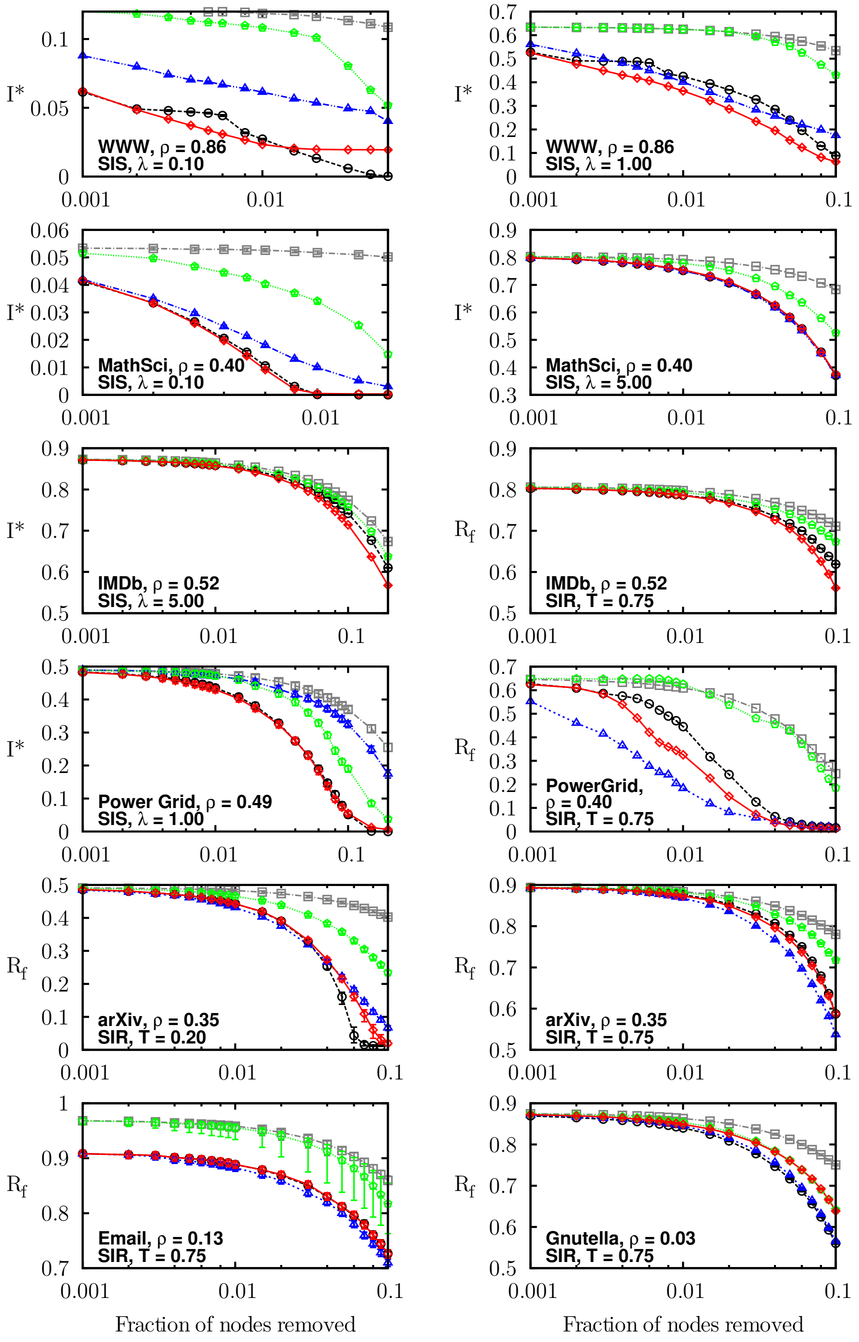


FIGURE B.6 – Efficiency of the immunization methods against SIS and SIR epidemics on several networks. Nodes are removed in decreasing order of their score according to each method: coreness (green pentagons), degree (black circles), betweenness centrality (blue triangles) and memberships (red diamonds) to measure efficiency by the decrease of I^* or R_f. The size of the epidemics for random removal of nodes (gray squares) is added for comparison. Error bars have been omitted for clarity of the SIR results on the Power Grid, but are shown in the SI.

B.4.3 Results on networks of diverse nature

In this section, we highlight different behaviours observed in social, technological and communication networks using 7 other datasets (full results for the 17 datasets are available in the SI): subset of the World Wide Web (WWW) [15], MathSciNet co-authorship network (MathSci) [155], Western States Power Grid of the United States (Power Grid) [197], Internet Movie Database since 2000 (IMDb) [88], cond-mat arXiv co-authorship network (arXiv) [153], e-mail interchanges between members of the University Rovira i Virgili (Email) [83] and Gnutella peer-to-peer network (Gnutella) [109].

The results for the WWW, MathSci and IMDb networks further support our previous conclusions, with the exception that membership-based immunization performs surprisingly better than the degree-based variant even near the epidemic threshold of the network (see WWW and MathSci). The betweenness-centrality-based immunization was not tested on IMDb because of computational constraints (its computation required over 800 hours with our available ressources and a standard algorithm [30]), which illustrates a significant limit of this measure. Approximations could have been used [114], but the intricate (and mostly unknown) relationship between the efficiency of the measure and the accuracy of the approximation would have only caused additional uncertainties.

The results presented for the Power Grid network illustrate a fundamental difference between the SIS and the SIR dynamics: while we are interested in the fraction of the network sustaining an established epidemic in SIS, it is the fraction of nodes invaded by a new disease that is relevant in SIR. In fact, the structure of the Power Grid, a chain of small, easily disconnected modules, enhances the qualitative discrepancy between the epidemic influence of nodes subjected to these two dynamics. For the SIS dynamics, the membership-based intervention is the most efficient because it weakens all modules, limiting the prevalence of the disease. In distinction, targeting through betweenness centrality merely separates the modules, so that they indiviually remain infected. For the SIR dynamics, separating the modules is the best approach as it directly stops the infection from spreading; while weakened – but connected – modules still provide pathways. This effect is a direct consequence of the particular structure of the Power Grid and is insignificant on other networks.

Finally, the last set of results, on arXiv, Email and Gnutella, present the effect of the community density ρ on the performance of membership-based immunization. For very small ρ, the paths within communities do not qualitatively differ from the links bridging neighborhoods in their effect on the disease propagation. This targeting method is therefore expected to converge toward degree-based immunization if m and k are strongly correlated. However, as most tested networks had fairly dense communities, $\rho \geq 0.3$, the relevance of memberships should not be understated.

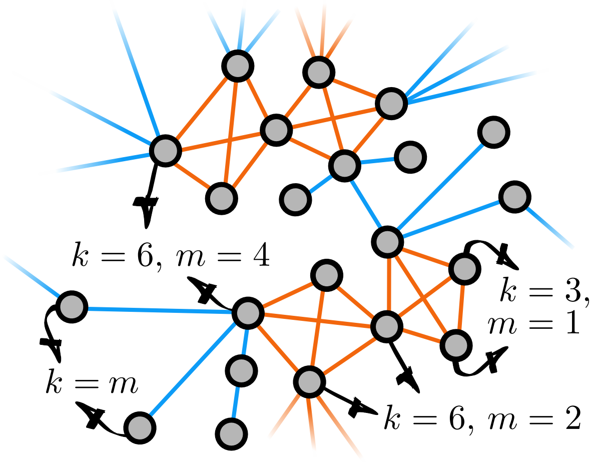


FIGURE B.7 – Synthetic networks with tunable community structure. Orange links belong to motifs of size $M = 4$, and single links are shown in blue. The degree k and membership m of a few selected nodes are indicated. They belong to $i = (k - m)/(M - 2)$ motifs and have $j = [(M - 1)m - k]/(M - 2)$ single links.

B.4.4 Investigation of the epidemic regimes transition

The results of the previous sections suggest that local information (i.e., degree, membership) is often sufficient for a nearly optimal global immunization. More precisely, we found these two methods to outperform or to be as efficient as the betweenness centrality (the global method used for comparison) in 62 of the 68 studied scenarios (i.e., 17 networks / 2 dynamics / 2 transmissibility regimes). This implies that membership (e.g., on PGP), degree (e.g., Gnutella) or both (e.g. MathSci) lead to an immunization at least as efficient as global methods while having the noteworthy advantage of requiring much less information and of being less sensitive to incomplete information. This section focuses on the conditions guiding the choice between the degree-based or the membership-based immunization strategy. In this respect, Figs. 4 and 6 provide a useful hindsight: the membership-based strategy is more efficient than the degree-based one when transmissibility is high and/or when communities are dense. To further our understanding and test this hypothesis, we introduce a random network model featuring a community structure, and exactly solve its final state (R_f) under SIR dynamics using generating functions.

Our model is a slightly modified version of the configuration model [147, 138] where nodes are connected either through single links or through motifs (see Fig. B.7 for an example). Motifs are used to simulate the effect of a community structure, that is the redundancy of the neighbourhoods of nodes. Our motifs are composed of M nodes, all connected to each other, and a node belongs to i motifs and has j single links with probability $p(i, j)$. This node therefore has a degree (k) equal to $(M-1)i+j$ and a membership (m) equal to $i+j$. Networks are generated using a stub pairing scheme: a node belonging to i motifs and having j single links has i "motif stubs" and j "link stubs". Groups and single links are then formed by randomly choosing M motif stubs and 2 link stubs, respectively, and then by linking the corresponding nodes to one another. This last step is repeated until none of the motif and link stubs remains. The distribution $\{p(i,j)\}_{i,j\in\mathbb{N}}$ therefore defines a maximally random network ensemble, and the results obtained are averaged over this ensemble.

Extending previous work [6], we compute the expected value of R_{f} for the network ensemble just defined where nodes and links are randomly removed to simulate immunization and disease transmission (SIR dynamics), respectively. Full details are given in the SI. Using typical values for $\{p(i,j)\}$, our model illustrates and confirms our hypothesis by clearly showing in Fig. B.8 a transition of efficiency between the degree-based and the membership-based immunization strategy. Initially less efficient when the transmissibility is low (i.e., higher threshold, lower value of R_{f}), membership progressively outperforms degree as the transmissibility increases. As mentionned above, for lower values of T, the best option is therefore to immunize the hubs (high k) to shift the degree distribution towards lower degrees. For higher values of T, targeting structural hubs (high m) that act as bridges between "independent" neighbourhoods leads to a more efficient immunization as it reduces the number of paths between different regions of the network. Note that we do not explicitly model the effect of community density. This could have been done by letting links exist independently with a given probability η. This is however identical to letting the disease propagate with probability ηT. Thus, the value of T in Fig. B.8 is related to the density of the communities, and our conclusions can therefore be extended to the cases of low/high community densities.

B.5 Discussions

One of the main contributions of this work is to offer a formal definition of the epidemic influence of nodes, i.e. the effect of its removal on I^* of R_{f}, which is open to diverse methods of approximation. Our results confirm that standard measures such as the degree or betweenness centrality are *not always* the best indicators of a node's influence. Moreover, we have highlighted that the coreness, which has recently been proposed as an indicator of nodes' influence [104], offers poor performances. This has brought us to distinguish between individual risk and global influence. We have also illustrated how a universal approach is still wanting, since different networks and different diseases require different methods of intervention.

Consequently, the fact that the numbers of links and/or communities to which a node belongs are excellent measure of its epidemic influence — at times better, at times equivalent, but never much worse than global centrality measures — is a particularly important result. The fact that they both are *local measures* is especially relevant considering that we rarely

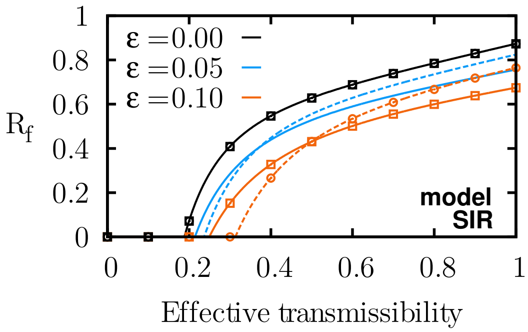


FIGURE B.8 – Results of local immunization methods on synthetic networks. Final sizes of SIR epidemics after immunization of various fractions ε of nodes on synthetic networks with $M = 4$ and an heterogeneous degree distribution (details in SI). Near the epidemic threshold, targeting by degree (dotted curves) is the better choice whereas targeting by memberships (solid curve) should be preferred for higher transmissibility. Monte Carlo simulations were also performed to validate the formalism and indicated on the curves (the case $\varepsilon = 0.05$ is omitted not to clutter the graph) with circles (targeting by degree) and squares (targeting by membership).

have access to the exact network structure of a system, either because it is simply too large (WWW), too dynamic (email networks) or because the links themselves are ill-defined (social networks). Not only are local measures computable from a limited subset of a network (which is often the only available information), but a coarse-grained measure like membership is even more interesting as it is easier to estimate than a node's actual degree. For instance, consider how much simpler it is to enumerate your social groups (work, family, etc.) than the totality of your acquaintances.

Finally, the existence of a transition between two epidemic regimes with different characteristic scales may well be the single most important conclusion of this work. In the first regime, for low transmissibility and sparse communities, the microscopic structural features (i.e. node connectivity or degree) offer the most relevant information; while for higher transmissibility and denser communities, mesoscopic features (i.e node communities or membership) appear more relevant. We expect to see an equivalent transition between any pair of measures which oppose the micro and meso scales (e.g. different range-limited measures of centrality [61]).

Based on our empirical and analytical results, we thus propose a simple procedure on how to judge which local measure can be expected to yield the best results in a given situation. From the available subset of a given network:

1. Obtain the degree distribution to estimate the transmissibility of the disease in relation to the epidemic threshold λ_c [93] or T_c [136].

2. If easily transmissible ($\lambda \gg \lambda_c$ or $T \gg T_c$), evaluate the network's community structure; otherwise, go to 4.
3. If the community density is high ($\rho \gtrsim 0.3$), immunize nodes according to their memberships; otherwise, go to 4.
4. For a transmissibility near the epidemic threshold, or for sparse communities (low ρ), immunize according to the degree of the nodes.

The analytical and numerical frameworks used in this work are expected to guide immunization efforts toward simpler, more precise and efficient strategies. Likewise, the introduction of a node influence classification scheme opens a new avenue for finding better local estimates of a node's role in the global state of its system.

B.6 Materials and methods

Betweenness centrality

For all pairs (a, b) of nodes excluding i, list the $n_{a,b}$ shortest paths between a and b. Let $n_{a,b}(i)$ be the number of these paths containing i. The betweenness centrality b_i of node i is then given by:

$$b_i = \sum_{(a,b)} \frac{n_{a,b}(i)}{n_{a,b}} \ . \tag{B.1}$$

Coreness

The coreness of node i is the highest integer c_i such that the node is part of the set of all nodes with at least c_i links within the set.

Community detection

Two links, e_{ij} and e_{ik}, from a given node i, are said to belong to the same community if their Jaccard coefficient $J(e_{ij}, e_{ik})$ (similarity measure) is above a given threshold J_c :

$$J\left(e_{ij}, e_{ik}\right) = \frac{n_+(j) \cap n_+(i)}{n_+(j) \cup n_+(i)} > J_c \ , \tag{B.2}$$

where $n_+(u)$ is the set containing the neighbors of u including u.

Community density

The density ρ_i of a community i of $n_i > 2$ nodes and d_i links is the proportion of the possible redundant links that do exist; i.e., the fraction of existing links excluding the minimal $n_i - 1$ links that are needed for this community to be connected:

$$\rho_i = \frac{d_i - (n_i - 1)}{\frac{n_i(n_i-1)}{2} - (n_i - 1)} \ . \tag{B.3}$$

The community density ρ is then calculated according to

$$\rho = \frac{1}{D} \sum_i d_i \rho_i \;, \tag{B.4}$$

where D is the total number of links not belonging to single link communities, for which $\rho_i = 0$ [2].

Immunization

To perform the immunization of a fraction ε of the network according to a certain measure Γ, we remove the εN nodes with the highest Γ. When a choice must be made (nodes with equal Γ), all decisions are taken randomly and individually for each simulated epidemics.

Monte Carlo simulations

To investigate the fraction of a network which can *sustain* an epidemics, SIS simulations start with all nodes in an infectious state and are then relaxed until an equilibrium is reached. To investigate the mean fraction of a network which a disease can *invade*, SIR simulations start with a single randomly chosen infectious node and run until there are no more infectious nodes. Results shown in the figures are obtained by averaging over the outcome of several numerical simulations until the minimal possible standard deviation (limited by network structure and finite size) is obtained. For the SIR dynamics, only the simulations leading to large-scale epidemics (at least 1% of the nodes) were considered. The complete procedure is given in the SI.

B.7 Appendix: Supplementary discussions on methods

Local vs global measures

We differentiate between these two types of measures by the information required to compute them. If this information (per node) is independent of total system size, the measure is considered local; whereas a global measure requires information scaling with system size (often a complete description). For the four properties studied in this paper, we thus consider that:

1. degree is a local measure, as only the number of neighbours of a node is required;
2. membership is a local measure, as the chosen algorithm only requires the neighbourhood of one given node and that of its neighbours to estimate its membership number;
3. coreness is a global measure, as a node's coreness depends on the coreness of its neighbours which in turn depend on the coreness of their neighbours and so on;
4. betweenness centrality is a global measure, since it is calculated by considering the shortest paths between a given node and all of the other nodes in the network.

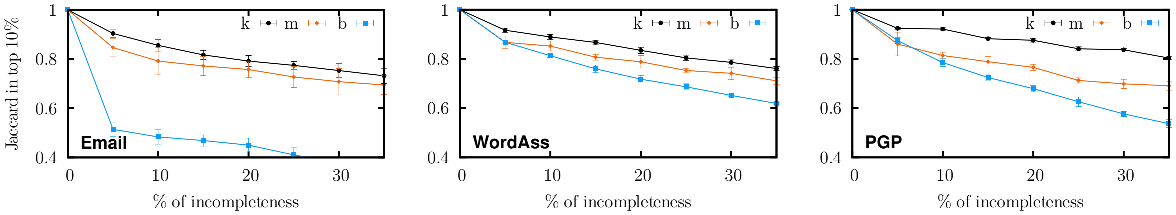


FIGURE B.9 – Robustness of measured from the micro (degree k), meso (memberships m) and macro (betweenness centrality b) scales when information on the network is removed. The robustness is here measured by comparing (with a Jaccard coefficient) the ensemble of nodes identified as being in the top 10% of nodes when a certain fraction of links are randomly removed (horizontal axis) as opposed to the ensemble obtained by considering the complete data.

For obvious reasons, *local measures are less sensitive to incomplete or incorrect information*. Adding, removing or rewiring a link only affects the degree or membership of nodes directly in the neighbourhood of the modification; whereas the same alterations can potentially affect the coreness or betweenness centrality of nodes anywhere in the network through cascading effects. Consequently, measures based on shorter-range information are always more robust to missing information, and sometimes quite significantly, as seen on Suppl. Fig. B.9.

B.7.1 Community detection

As mentioned above the *link clustering algorithm* of Ahn *et al.* was chosen in part because it can perform well (and at times even better) using local information instead of the entire network [2]. While we always partitioned the network globally, by setting a resolution threshold, the identification of structural hubs is very robust to this global threshold (see Suppl. Fig. B.10). More importantly, this algorithm groups links stemming from a given node in a community based on the similarity of the two neighbourhoods reached through them. Hence, it evaluates the redundancy in second neighbourhoods (how many of my second neighbours are neighbours of more than one of my neighbours?). This redundancy (or overlap) can then serve as an appropriate measure to gauge the major impact of community structure on an epidemic process, namely the loss of potential new infections due to clustering. The link clustering algorithm therefore provides a well-defined method to quantify this loss.

B.7.2 Supplementary simulation details

SIS

All nodes are initially infectious and we relax the system by iterating a discrete time propagation simulation using time step Δt chosen such that $\alpha\Delta t$ and $\beta\Delta t$ are less than 10^{-3}:

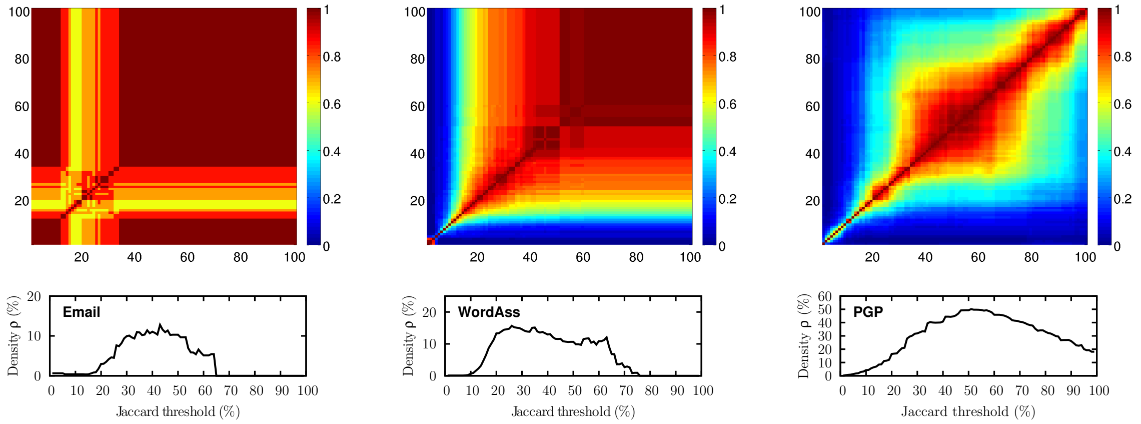


FIGURE B.10 – Robustness of our ability to identify structural hubs (1% of nodes with the most memberships) as the global resolution varies (the final partition is chosen to maximize the community density shown at the bottom). The color map represents the Jaccard coefficient, i.e. the similarity of ensembles, as measured between the structural hubs identified with two different resolution parameter. These ensembles are always very similar when avoiding extreme resolutions (e.g. low density). Note that in the color map, yellow corresponds to a Jaccard coefficient of 0.6 which implies that 75% of the same structural hubs were found by both partitions.

i. at each Δt, every susceptible neighbour S of every infectious individual I is infected with probability $\alpha\Delta t$;
ii. at each Δt every infectious individual I recovers with probability $\beta\Delta t$;

the steady-state is averaged over multiple independent simulations to minimize the standard deviation (due to network structure and finite size).

SIR

A single node is randomly infected and the following stochastic process is iterated until no infectious nodes remain:

i. I nodes infect each of their S neighbours with probability T and then recover.

The final state, considering only epidemics larger than 1% of the system size, is averaged over multiple independent simulations to minimize the standard deviation (due to network structure and finite size).

B.8 Appendix: Theoretical modelling

The conclusion drawn in the main text were validated using synthetic networks. In this Section, we present how these synthetic networks are generated. Furthermore, we describe the mathematical framework used to calculate the final outcome of the SIR dynamics on these networks. Finally, we give the parameters used for the main results.

B.8.1 Synthetic networks

The synthetic networks considered are a clustered and multitype generalisation of the Configuration Model [6]. In these networks, nodes are connected either through single links or through motifs (see Fig. 7 in the main text for an example). Motifs are composed of M nodes which are all connected to one another, and a node belongs to i motifs and has j single links with probability $p(i,j)$. This node therefore has a degree (k) equal to $(M-1)i+j$ and a membership (m) equal to $i+j$.

Networks are generated using a stub pairing scheme: a node belonging to i motifs and having j single links has i "motif stubs" and j "link stubs". Groups and single links are then formed by randomly choosing M motif stubs and 2 link stubs, respectively, and then by linking the corresponding nodes to one another. This last step is repeated until none of the motif and link stubs remain. The distribution $\{p(i,j)\}_{i,j\in\mathbb{N}}$ therefore defines a random network ensemble, and the results obtained in this Section are averaged over this ensemble.

B.8.2 Mathematical formalism

As there exists a mapping—under simple assumptions—between the SIR dynamics and bond percolation on networks [102, 136]. To calculate the outcome of the SIR dynamics on the networks just described, we use a previously published formalism [6] in which we add the possibility for nodes to exist with a given probability (i.e., site percolation) to simulate immunization strategies. We only give a short outline of this theoretical model as a general and more formal description will be the subject of a subsequent publication.

For each pair (i,j) such that $p(i,j) \neq 0$ we assign a node type denoted by $\{i,j\}$ (the set of such pairs is noted $\mathcal{M}$). As the pair (i,j) is the only information available about the nodes, assigning one node type per pair allows us to simulate very detailed immunization strategies. Indeed, we define $q_{\{i,j\}}$ as the probability for a type-$\{i,j\}$ node to be immunized; the simulated immunization strategy is therefore encoded in the set of probabilities $\{q_{\{i,j\}}\}_{\{i,j\}\in\mathcal{M}}$. Also, as explained in the main text, the infectious agent propagates from an infected node to a susceptible neighbour with probability T. From a percolation point of view, $1-q_{\{i,j\}}$ is the occupation probability of type-$\{i,j\}$ sites (nodes) and T is the occupation probability of bonds (links).

Solving site/bond percolation in motifs

The mathematical formalism that we have developed relies on probability generating functions (PGFs) and therefore implicitly requires the networks under consideration to have a tree-like structure. As the networks we consider contain motifs, which clearly do not comply with that assumption, we need to solve the bond and site percolation within motifs beforehand.

As previously shown [7], the bond percolation outcome—the distribution of the number of nodes that can be reached by following links from an initial node—can be exactly obtained by iterating a set of simple equations. We denote $\boldsymbol{n}$ the $|\mathcal{M}|$-tuple whose elements[3] $n_{\{i,j\}}$ correspond to the number of nodes of each type (i.e., there are $n_{\{i,j\}}$ type-$\{i,j\}$ nodes). For the remaining, each boldfaced variable will correspond to such $|\mathcal{M}|$-tuple.

Let us define $Q_{\{i,j\}}(\boldsymbol{l}|\boldsymbol{n})$ as the probability to find a component of $\boldsymbol{l}$ nodes in a motif of size (and composition) $\boldsymbol{n}$ from an initial type-$\{i,j\}$ node. Although nodes are initially all connected to one another in motifs, we are interested in the number of nodes (and their type) that can be reached from an initial node when links are followed with a probability T (bond percolation). Following previous work [7], $Q_{\{i,j\}}(\boldsymbol{l}|\boldsymbol{n})$ is obtained by iterating

$$Q_{\{i,j\}}(\boldsymbol{l}|\boldsymbol{n}) = Q_{\{i,j\}}(\boldsymbol{l}|\boldsymbol{l}) \prod_{\{i',j'\}\in\mathcal{M}} \binom{n_{\{i',j'\}} - \delta_{ii'}\delta_{jj'}}{l_{\{i',j'\}} - \delta_{ii'}\delta_{jj'}} \prod_{\{i'',j''\}\in\mathcal{M}} (1-T)^{n_{\{i',j'\}}(n_{\{i'',j''\}} - l_{\{i'',j''\}})} \tag{B.5a}$$

$$Q_{\{i,j\}}(\boldsymbol{l}|\boldsymbol{l}) = 1 - \sum_{\boldsymbol{m}<\boldsymbol{l}} Q_{\{i,j\}}(\boldsymbol{m}|\boldsymbol{l}) \tag{B.5b}$$

from the initial condition $Q_{\{i,j\}}(\boldsymbol{\delta_{\{i,j\}}}|\boldsymbol{\delta_{\{i,j\}}})$, where δ_{ij} is the Kronecker delta, and where $\boldsymbol{\delta_{\{i,j\}}}$ is an $|\mathcal{M}|$-tuple whose elements are all equal to 0 except for the $\{i,j\}$-th one that is equal to one. The sum in Eq. (B.5b) is over all $\boldsymbol{m}$ such that $m_{\{i,j\}} \leq l_{\{i,j\}}$ for every node type $\{i,j\}$ but with the additional constraint that $\boldsymbol{m} \neq \boldsymbol{l}$. The initial condition simply states that the probability of finding a component of one type-$\{i,j\}$ node from a type-$\{i,j\}$ node in a motif containing only one type-$\{i,j\}$ node is one (the initial node is always included in the size of the component). Then, Eqs. (B.5) iteratively increase the size of the motif and compute the size distribution along the way until the complete distribution for a motif of the desired size (and composition) is obtained.

Should we be interested in studying bond percolation on motifs solely, we would keep the distribution $\{Q_{\{i,j\}}(\boldsymbol{l}|\boldsymbol{n})\}$ for a given size $\boldsymbol{n}$, and discard the distributions for motifs of intermediate size obtained while iterating Eqs. (B.5). These intermediate distributions can however be used to *exactly* predict the distribution of the number of nodes that can be reached by following links from an initial node in motifs where links *and* nodes exist with given probabilities (bond and site percolation). Indeed, as each node exists independently with a given probability, the probability for a motif of original size $\boldsymbol{n}$ to be of size $\boldsymbol{m}$ after the random removal of its nodes is simply

$$W_{\{i,j\}}(\boldsymbol{m}|\boldsymbol{n}) = \prod_{\{i',j'\}\in\mathcal{M}} \binom{n_{\{i',j'\}} - \delta_{ii'}\delta_{jj'}}{m_{\{i',j'\}} - \delta_{ii'}\delta_{jj'}} \Big[1 - q_{\{i',j'\}}\Big]^{m_{\{i',j'\}} - \delta_{ii'}\delta_{jj'}} \Big[q_{\{i',j'\}}\Big]^{n_{\{i',j'\}} - m_{\{i',j'\}}} , \tag{B.6}$$

[3] $|\mathcal{M}|$ is the number of elements (i.e., the cardinality) of the set $\mathcal{M}$, hence the number of node types.

where we assume that the initial type-$\{i,j\}$ exists. Then, the probability for a type-$\{i,j\}$ node to lead to a component of size $\boldsymbol{l}$ in a motif of original size $\boldsymbol{n}$ but whose links and nodes have been randomly removed is simply

$$P_{\{i,j\}}(\boldsymbol{l}|\boldsymbol{n}) = \sum_{\boldsymbol{m}=\delta_{\{i,j\}}}^{\boldsymbol{n}} Q_{\{i,j\}}(\boldsymbol{l}|\boldsymbol{m}) W_{\{i,j\}}(\boldsymbol{m}|\boldsymbol{n}) \; . \tag{B.7}$$

The site and bond percolation can therefore be exactly solved for motifs by iterating Eqs. (B.5)–(B.7). As expected, if applied to the simplest motifs, i.e. links, the type-$\{i,j\}$ node at the other end of the link will be reached with probability $T(1-q_{\{i,j\}})$ and will not be reached with probability $1-T(1-q_{\{i,j\}})$.

As motifs and single links are built by randomly matching stubs, the probability for a type-$\{i,j\}$ node to appear in a motif [a link] is proportional to $[ip(i,j)]$ $[jp(i,j)]$. The probability for a motif to be composed by $\boldsymbol{n}$ nodes is given by the multinomial distribution

$$R^{(\mathrm{m})}(\boldsymbol{n}) = \frac{M!}{\left[\sum_{\{i,j\}\in\mathcal{M}} ip(i,j)\right]^{M}} \prod_{\{i,j\}\in\mathcal{M}} \frac{\left[ip(i,j)\right]^{n_{\{i,j\}}}}{n_{\{i,j\}}!} \; . \tag{B.8a}$$

The same applies for the composition of links

$$R^{(\mathrm{l})}(\boldsymbol{n}) = \frac{2!}{\left[\sum_{\{i,j\}\in\mathcal{M}} jp(i,j)\right]^{2}} \prod_{\{i,j\}\in\mathcal{M}} \frac{\left[jp(i,j)\right]^{n_{\{i,j\}}}}{n_{\{i,j\}}!} \; . \tag{B.8b}$$

Finally, for the purpose of calculating the outcome of the bond and site percolation on the synthetic networks, let us define the two following generating functions:

$$\theta^{(\mathrm{m})}_{\{i,j\}}(\boldsymbol{x}) = \sum_{\boldsymbol{n}} \frac{n_{\{i,j\}} R^{(\mathrm{m})}(\boldsymbol{n})}{\sum_{\boldsymbol{n}'} n'_{\{i,j\}} R^{(\mathrm{m})}(\boldsymbol{n}')} \left[\sum_{\boldsymbol{l}=\delta_{\{i,j\}}}^{\boldsymbol{n}} P^{(\mathrm{m})}_{\{i,j\}}(\boldsymbol{l}|\boldsymbol{n}) \prod_{\{i',j'\}\in\mathcal{M}} \left[x_{\{i',j'\}}\right]^{l_{\{i',j'\}}-\delta_{ii'}\delta_{jj'}} \right] \tag{B.9a}$$

$$\theta^{(\mathrm{l})}_{\{i,j\}}(\boldsymbol{y}) = \sum_{\boldsymbol{n}} \frac{n_{\{i,j\}} R^{(\mathrm{l})}(\boldsymbol{n})}{\sum_{\boldsymbol{n}'} n'_{\{i,j\}} R^{(\mathrm{l})}(\boldsymbol{n}')} \left[\sum_{\boldsymbol{l}=\delta_{\{i,j\}}}^{\boldsymbol{n}} P^{(\mathrm{l})}_{\{i,j\}}(\boldsymbol{l}|\boldsymbol{n}) \prod_{\{i',j'\}\in\mathcal{M}} \left[y_{\{i',j'\}}\right]^{l_{\{i',j'\}}-\delta_{ii'}\delta_{jj'}} \right] \tag{B.9b}$$

where the superscript “m” (resp. “l”) indicate that the quantities have been solved for motifs (resp. links). In other words, the function $\theta^{(\mathrm{m})}_{\{i,j\}}(\boldsymbol{x})$ generates the probability distribution for the number of nodes of each type that can be reached from a type-$\{i,j\}$ node in a random motif [i.e., whose composition is averaged over $R^{(\mathrm{m})}(\boldsymbol{n})$]. Specifically, the coefficient in front of $x^{s}_{\{i',j'\}}$ in Eq. (B.9a) is the probability of reaching s type-$\{i',j'\}$ nodes from a type-$\{i,j\}$ node in a random motif. The same applies to $\theta^{(\mathrm{l})}_{\{i,j\}}(\boldsymbol{y})$.

Calculating the average fate of an outbreak

We are now in a position to solve the bond and site percolation on the synthetic networks defined previously. For the purpose of the present study, we are interested in the quantity

R_f: the relative size of the extensive (giant) component. To highlight the nontrivial effect of immunization[4], R_f is expressed in terms of the fraction of the *existing* nodes (i.e., $1-\varepsilon$) that are part of the giant component.

It is convenient to introduce the following function

$$g_{\{i,j\}}(\boldsymbol{x},\boldsymbol{y}) = \left[\theta^{(\mathrm{m})}_{\{i,j\}}(\boldsymbol{x})\right]^{i}\left[\theta^{(\mathrm{l})}_{\{i,j\}}(\boldsymbol{y})\right]^{j} \tag{B.10}$$

generating the distribution of the number of nodes of each type that are in the immediate neighbourhood of a type-$\{i,j\}$ node. The *immediate neighbourhood* refers to the nodes to which the type-$\{i,j\}$ node is connected either via its single links or via its motifs. Similarly, we define

$$f^{(\mathrm{m})}_{\{i,j\}}(\boldsymbol{x},\boldsymbol{y}) = \left[\theta^{(\mathrm{m})}_{\{i,j\}}(\boldsymbol{x})\right]^{i-1}\left[\theta^{(\mathrm{l})}_{\{i,j\}}(\boldsymbol{y})\right]^{j} \tag{B.11a}$$

$$f^{(\mathrm{l})}_{\{i,j\}}(\boldsymbol{x},\boldsymbol{y}) = \left[\theta^{(\mathrm{m})}_{\{i,j\}}(\boldsymbol{x})\right]^{i}\left[\theta^{(\mathrm{l})}_{\{i,j\}}(\boldsymbol{y})\right]^{j-1} \tag{B.11b}$$

generating the distribution of the number of nodes of each type that are in the immediate neighbourhood of a type-$\{i,j\}$ node that has been reached by either one of its single links or one of the motifs it is a part of (if applicable). In other words, these functions generate the excess degree distribution.

To calculate R_f, let us define $a_{\{i,j\}}$ as the probability that a link to a type-$\{i,j\}$ node *does not* lead to the giant component. Similarly, we define $b_{\{i,j\}}$ as the probability that a type-$\{i,j\}$ node reached through a motif does not lead to the giant component. Due to the effective tree-like structure of the networks—recall that the outcome of percolation on the motifs has already been solved—$a_{\{i,j\}}$ and $b_{\{i,j\}}$ must satisfy the following self-consistency relations

$$a_{\{i,j\}} = f^{(\mathrm{m})}_{\{i,j\}}(\boldsymbol{a},\boldsymbol{b}) \tag{B.12a}$$

$$b_{\{i,j\}} = f^{(\mathrm{l})}_{\{i,j\}}(\boldsymbol{a},\boldsymbol{b}) \ . \tag{B.12b}$$

Put simply, these equations state that if a type-$\{i,j\}$ node reached from either a link or a motif does not lead to the giant component, then neither should the nodes that can be reached from it. The probability that a type-$\{i,j\}$ node *is* part of the giant component is then $1 - g_{\{i,j\}}(\boldsymbol{a},\boldsymbol{b})$. The probability that a randomly existing node is part of the giant component—which corresponds to its size as well—is therefore

$$R_\mathrm{f} = \sum_{\{i,j\}\in\mathcal{M}} \frac{(1-q_{\{i,j\}})p(i,j)\left[1-g_{\{i,j\}}(\boldsymbol{a},\boldsymbol{b})\right]}{\sum_{\{i',j'\}\in\mathcal{M}}(1-q_{\{i',j'\}})p(i',j')} \ . \tag{B.13}$$

[4] The relative size of the giant component cannot exceed $1-\varepsilon$ on networks for which a fraction ε of the nodes has been removed. This reduction in size obviously occurs during any immunization strategy and, for comparison purposes, must be taken into account.

$\{i,j\}$	$p(i,j)$	$k_{\{i,j\}}$	$m_{\{i,j\}}$
$\{0,1\}$	0.43930	1	1
$\{0,2\}$	0.13179	2	2
$\{0,9\}$	0.00712	9	9
$\{1,3\}$	0.25831	6	4
$\{1,6\}$	0.04982	9	7
$\{2,3\}$	0.02325	9	5
$\{3,0\}$	0.09041	9	3

TABLE B.2 – Distribution $\{p(i,j)\}_{i,j\in\mathbb{N}}$ used for the synthetic networks discussed in the main text. The degree and the membership of each node type is computed according to $k_{\{i,j\}} = (M-1)i + j$ and $m_{\{i,j\}} = i + j$, respectively, with $M = 4$.

The theoretical predictions (lines) on Fig. 8 in the main text were obtained by solving Eqs. (B.5)–(B.13) for various values of T and $\{q_{\{i,j\}}\}_{\{i,j\}\in\mathcal{M}}$. Comparison with results obtained from numerical simulations (symbols) confirms the validity of our theoretical model.

Parameters used for theoretical calculations

Table B.2 shows the distribution $\{p(i,j)\}_{i,j\in\mathbb{N}}$ used for the synthetic networks considered in the main text. It also gives the degree ($k_{\{i,j\}}$) and the membership ($m_{\{i,j\}}$) of each node type $\{i,j\} \in \mathcal{M}$. Motifs were composed of $M = 4$ nodes, and numerical simulation results (symbols on Fig. 8) were averaged over 5×10^5 realisations of networks of 2.5×10^5 nodes. For node types with $k_{\{i,j\}} > 2$, we let sets of $M-1$ links to either be part of cliques of M nodes or be single links in order to avoid unintended degree correlations [103]. For a given fraction ε of the nodes to immunize, we have

$$\sum_{\{i',j'\}\in\mathcal{M}} q_{\{i',j'\}} p(i',j') = \varepsilon \,. \tag{B.14}$$

The probabilities $q_{\{i,j\}}$ are chosen to satisfy this condition and in decreasing order of degree or membership.

B.9 Appendix: Supplementary results

The last sections of this Supplementary Information present a more complete view of the results obtained on empirical networks and are structured as follows. Each section covers one of the 17 datasets used in the study. Firstly, a brief discussion on the nature of each network is given, along with:

- the number of nodes (N), of links (L) and the degree distribution (k links per node);
- the maximal community density ρ and corresponding Jaccard threshold J_ρ.
- the maximal values of degree k, coreness c, betweenness centrality b and memberships m.

Secondly, correlations between degree, betweenness centrality, coreness and memberships are quantified using Spearman's rank correlation coefficient (defined below). We leave to the reader to observe how, given the correlation coefficient between memberships ranking and degree ranking, along with the mean community density, one can somewhat predict if the membership-based immunization will be more or less efficient than the degree-based version. Finally, the results of all immunization methods (random or on the four measures) are presented for SIS and SIR dynamics for a virulence (probability of disease transmission) close and far from the network's epidemic threshold.

B.9.1 Spearman's rank correlation coefficient

The Spearman's rank correlation coefficient quantifies the statistical dependence of two different orderings of the same set of items (nodes) on a scale of -1 (perfectly anti-correlated) to 1 (perfectly correlated) [183].

Consider x_i to be the rank of item i according to measure X, and y_i to be the rank of the same item according to a different measure Y. If for example, 10 items have the same score according to X and would otherwise be ranked from x_j to x_{j+9}, they are all given the rank $\left[\sum_{k=0}^{9} x_{j+k}\right]/10$. The Spearman's rank correlation coefficient $\sigma(X,Y)$ is then given by:

$$\sigma(X,Y) = \left[\sum_i (x_i - \bar{x})(y_i - \bar{y})\right] \Big/ \left[\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2\right]^{1/2} , \tag{B.15}$$

where $\bar{u}$ is the average rank according to measure U (the mean of $\{u_i\}$).

B.9.2 arXiv co-authorship

The cond-mat arXiv database uses articles published at `http://arxiv.org/archive/cond-mat` between April 1998 and February 2004. In this network, an article written by n co-authors contributes to a link of weight $(n-1)$ between every pair of authors. The unweighted network was obtained by deleting all links with a weight under the selected threshold of 0.1 [153].

TABLE B.3 – arXiv statistics

N	L	$k_{\max}$	$c_{\max}$	$b_{\max}$	$m_{\max}$	ρ
30561	125959	191	15	$6.9e+06$	127	0.35

TABLE B.4 – arXiv correlations

$\sigma(k,m)$	$\sigma(b,m)$	$\sigma(c,m)$	$\sigma(k,b)$	$\sigma(k,c)$	$\sigma(b,c)$
0.7766	0.7717	0.6639	0.7461	0.9411	0.5388

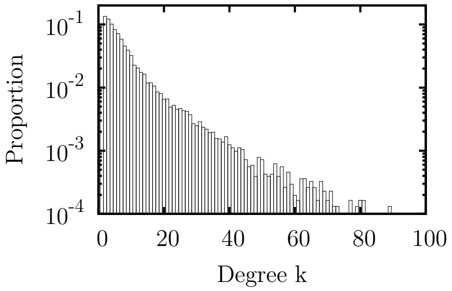


FIGURE B.11 – arXiv degree distribution

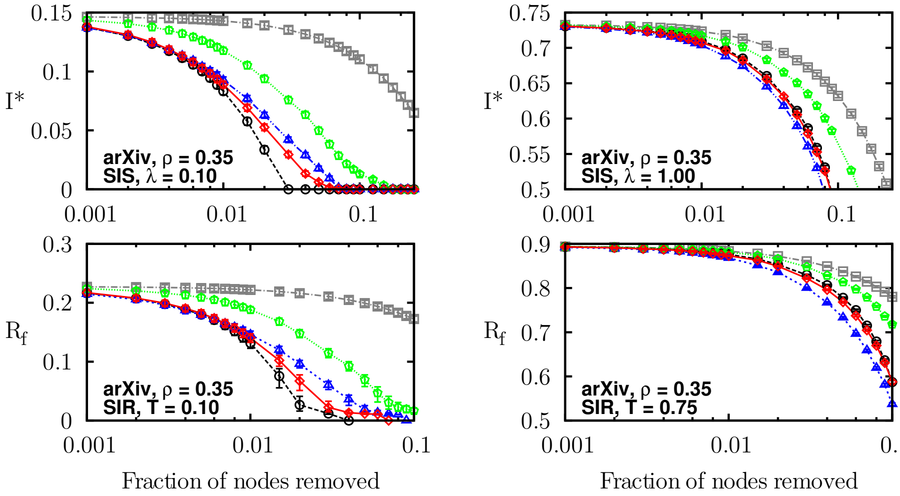


FIGURE B.12 – Intervention against epidemics on arXiv after different immunization: randomly (grey squares) and based on coreness (green pentagons), degree (black circles), betweenness centrality (blue triangles) or memberships (red diamonds).

B.9.3 Brightkite online social network

Brightkite was a location-based online social network where users could “check in” to the physical places they were visiting to connect with nearby friends. This datasets was obtained from a total of 4,491,143 check-ins over the period of Apr. 2008 - Oct. 2010 [41].

TABLE B.5 – Brightkite statistics

N	L	$k_{\max}$	$c_{\max}$	$b_{\max}$	$m_{\max}$	ρ
58228	214078	1134	52	$2e+08$	1118	0.55

TABLE B.6 – Brightkite correlations

$\sigma(k,m)$	$\sigma(b,m)$	$\sigma(c,m)$	$\sigma(k,b)$	$\sigma(k,c)$	$\sigma(b,c)$
0.9845	0.8919	0.9477	0.8822	0.9659	0.7767

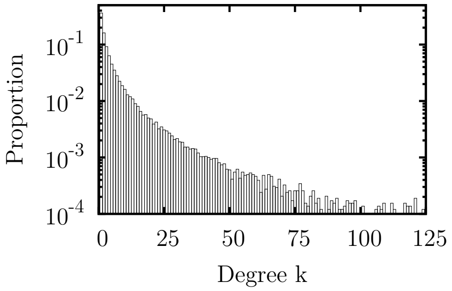


FIGURE B.13 – Brightkite degree distribution

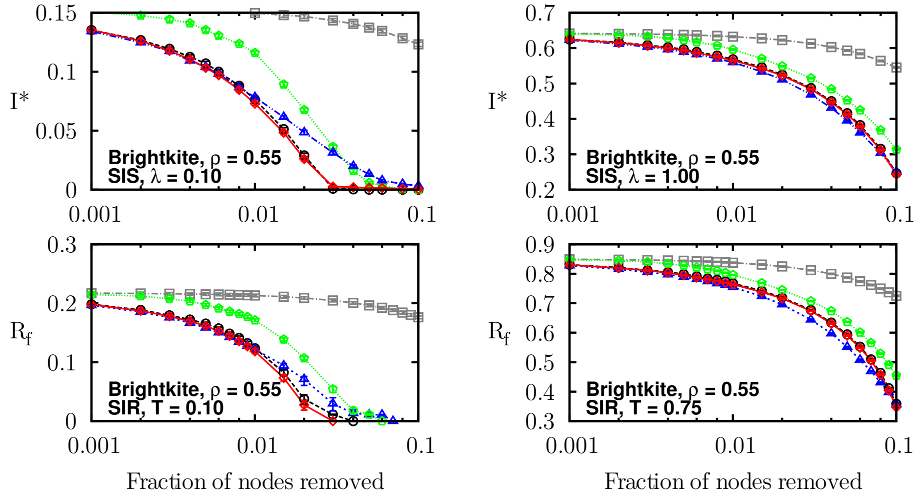


FIGURE B.14 – Intervention against epidemics on Brightkite after different immunization: randomly (grey squares) and based on coreness (green pentagons), degree (black circles), betweenness centrality (blue triangles) or memberships (red diamonds).

B.9.4 University email exchange

Network of email communication between accounts from the University Rovira i Virgili [83].

TABLE B.7 – Email statistics

N	L	$k_{\max}$	$c_{\max}$	$b_{\max}$	$m_{\max}$	ρ
1134	5143	1080	8	$6.1e+05$	929	0.13

TABLE B.8 – Email correlations

$\sigma(k,m)$	$\sigma(b,m)$	$\sigma(c,m)$	$\sigma(k,b)$	$\sigma(k,c)$	$\sigma(b,c)$
0.9900	0.9474	0.9560	0.9447	0.9613	0.8831

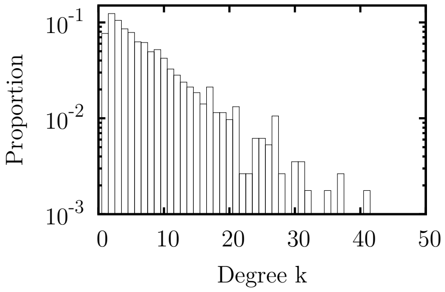


FIGURE B.15 – Email degree distribution

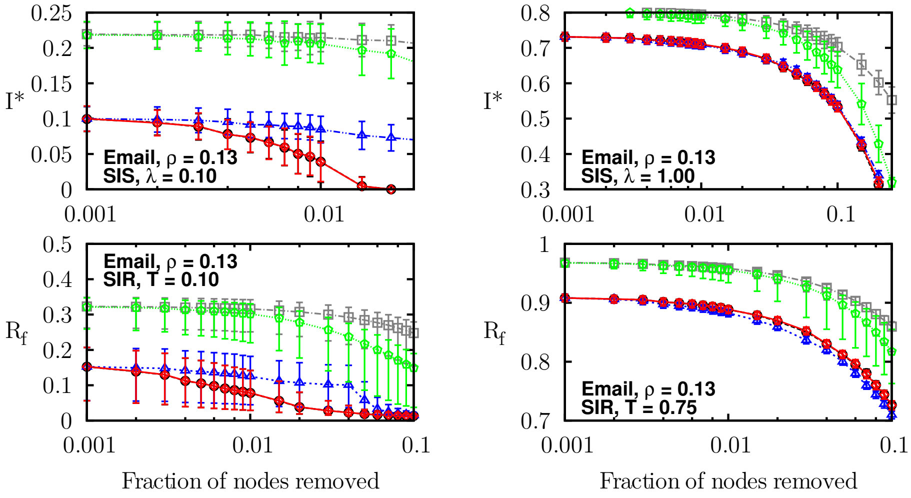


FIGURE B.16 – Intervention against epidemics on university email network after different immunization: randomly (grey squares) and based on coreness (green pentagons), degree (black circles), betweenness centrality (blue triangles) or memberships (red diamonds).

B.9.5 Enron email exchange

Network of email interchanges between all different Enron email addresses built from a dataset of around half million emails (made public by the Federal Energy Regulatory Commission) [106].

TABLE B.9 – Enron statistics

N	L	$k_{\max}$	$c_{\max}$	$b_{\max}$	$m_{\max}$	ρ
36692	183831	1383	43	$4.3e+07$	1306	0.61

TABLE B.10 – Enron correlations

$\sigma(k,m)$	$\sigma(b,m)$	$\sigma(c,m)$	$\sigma(k,b)$	$\sigma(k,c)$	$\sigma(b,c)$
0.9325	0.7567	0.9173	0.7585	0.9839	0.6862

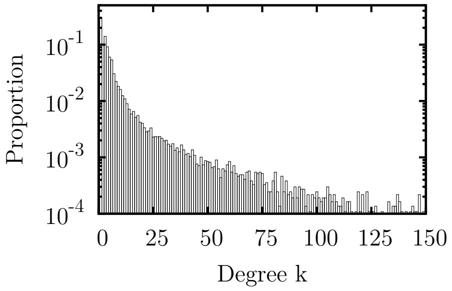


FIGURE B.17 – Enron degree distribution

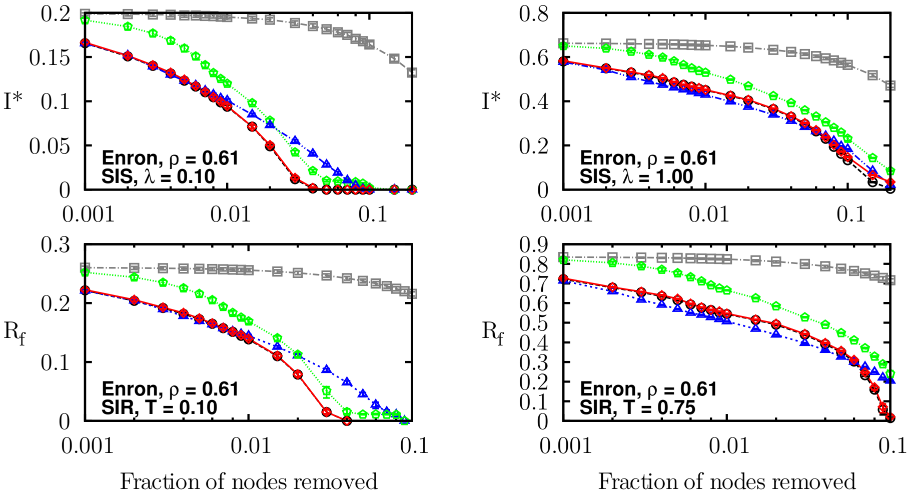


FIGURE B.18 – Intervention against epidemics on Enron email network after different immunization: randomly (grey squares) and based on coreness (green pentagons), degree (black circles), betweenness centrality (blue triangles) or memberships (red diamonds).

B.9.6 Gnutella peer-to-peer network

A snapshot of the Gnutella peer-to-peer network, where nodes are hosts and edges connections, from August 30th 2002. The data is originally directed (files taken *from* one host *to* another), but was made undirected for this work [166].

TABLE B.11 – Gnutella statistics

N	L	$k_{\max}$	$c_{\max}$	$b_{\max}$	$m_{\max}$	ρ
36682	88328	55	7	$5.3e+06$	52	0.03

TABLE B.12 – Gnutella correlations

$\sigma(k,m)$	$\sigma(b,m)$	$\sigma(c,m)$	$\sigma(k,b)$	$\sigma(k,c)$	$\sigma(b,c)$
0.9849	0.9848	0.9796	0.9925	0.9823	0.9743

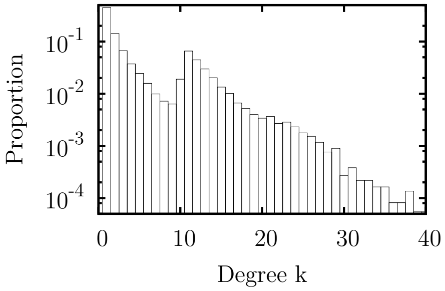


FIGURE B.19 – Gnutella degree distribution

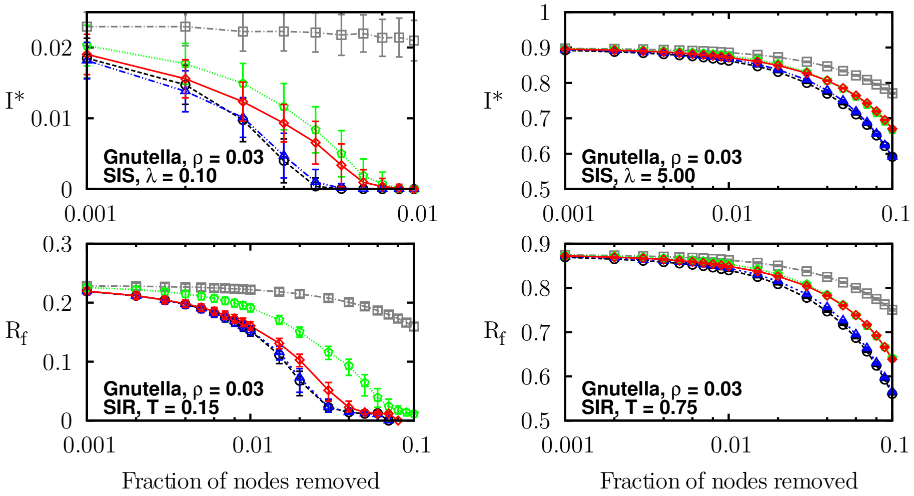


FIGURE B.20 – Intervention against epidemics on Gnutella network after different immunization: randomly (grey squares) and based on coreness (green pentagons), degree (black circles), betweenness centrality (blue triangles) or memberships (red diamonds).

B.9.7 Google weblinks

Directed network of hyperlinks between Google's webpages (considered undirected for this study) [154].

TABLE B.13 – Google statistics

N	L	$k_{\max}$	$c_{\max}$	$b_{\max}$	$m_{\max}$	ρ
15763	149456	11401	102	$9.0e+07$	2883	0.49

TABLE B.14 – Google correlations

$\sigma(k,m)$	$\sigma(b,m)$	$\sigma(c,m)$	$\sigma(k,b)$	$\sigma(k,c)$	$\sigma(b,c)$
0.8862	0.7941	0.8401	0.7735	0.9723	0.6995

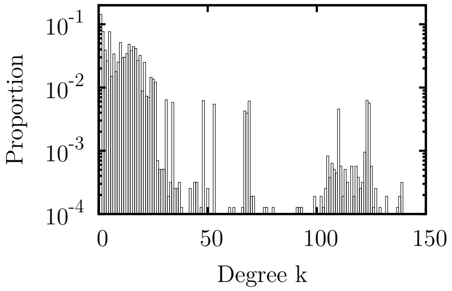


FIGURE B.21 – Google degree distribution

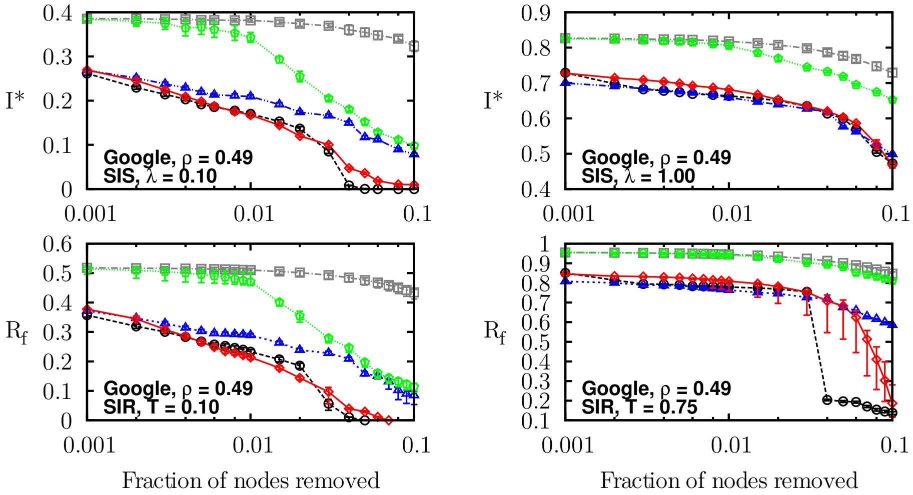


FIGURE B.22 – Intervention against epidemics on Google network after different immunization: randomly (grey squares) and based on coreness (green pentagons), degree (black circles), betweenness centrality (blue triangles) or memberships (red diamonds).

B.9.8 Gowalla social network

Gowalla is a location-based social networking website similar to Brightkite. This friendship network is undirected and composed from a total of 6,442,890 check-ins over the period of Feb. 2009 - Oct. 2010 [41].

TABLE B.15 – Gowalla statistics

N	L	$k_{\max}$	$c_{\max}$	$b_{\max}$	$m_{\max}$	ρ
196591	950327	14730	51	$6.3e+09$	14600	0.54

TABLE B.16 – Gowalla correlations

$\sigma(k,m)$	$\sigma(b,m)$	$\sigma(c,m)$	$\sigma(k,b)$	$\sigma(k,c)$	$\sigma(b,c)$
0.9792	0.8363	0.9514	0.8311	0.9724	0.7309

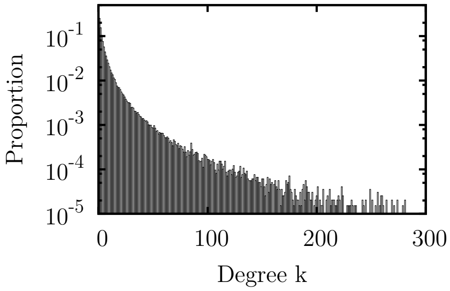


FIGURE B.23 – Gowalla degree distribution

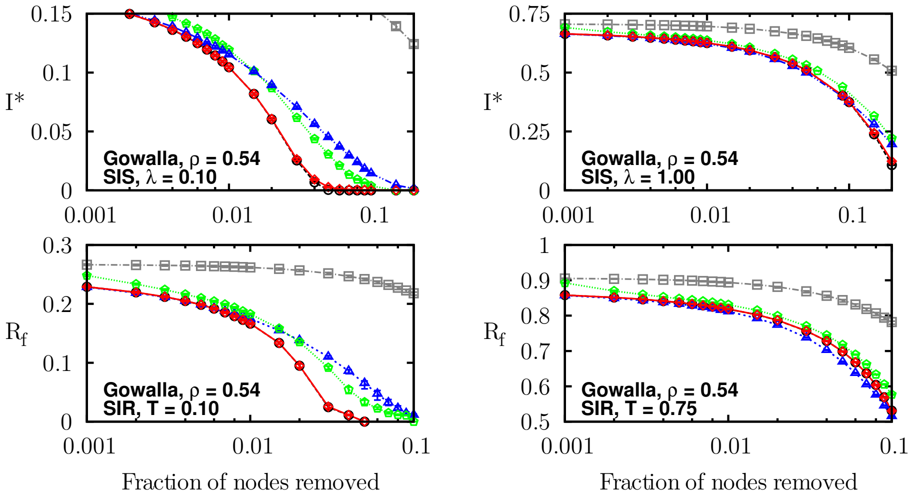


FIGURE B.24 – Intervention against epidemics on Gowalla network after different immunization: randomly (grey squares) and based on coreness (green pentagons), degree (black circles), betweenness centrality (blue triangles) or memberships (red diamonds).

B.9.9 Internet autonomous systems

This dataset is a symmetrized snapshot of the structure of the Internet at the level of autonomous systems, reconstructed from BGP tables posted at *archive.routeviews.org*. This snapshot was created by Mark Newman from data for July 22nd 2006 [88].

TABLE B.17 – Internet statistics

N	L	$k_{\max}$	$c_{\max}$	$b_{\max}$	$m_{\max}$	ρ
22963	48436	2390	25	$3.8e+07$	1710	$7e-4$

TABLE B.18 – Internet correlations

$\sigma(k,m)$	$\sigma(b,m)$	$\sigma(c,m)$	$\sigma(k,b)$	$\sigma(k,c)$	$\sigma(b,c)$
0.9857	0.7933	0.9469	0.7807	0.9631	0.7079

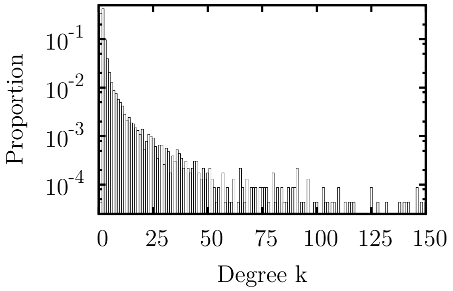


FIGURE B.25 – Internet degree distribution

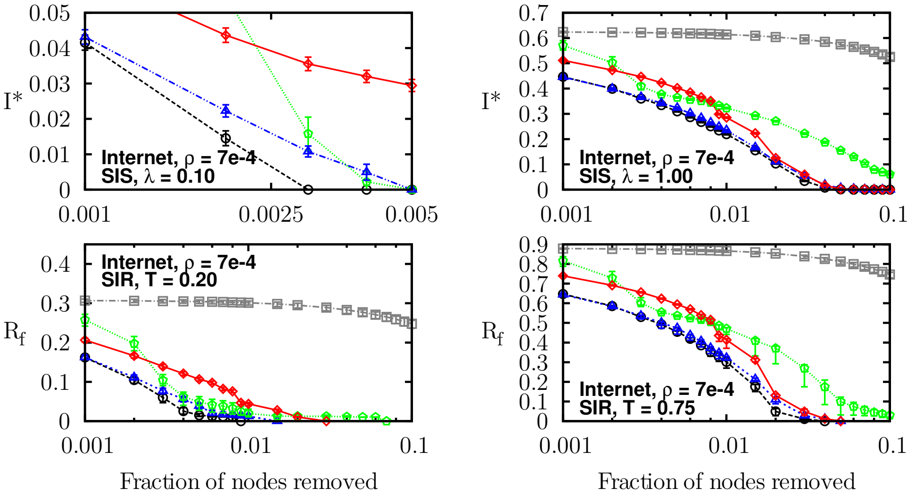


FIGURE B.26 – Intervention against epidemics on Internet network after different immunization: randomly (grey squares) and based on coreness (green pentagons), degree (black circles), betweenness centrality (blue triangles) or memberships (red diamonds).

B.9.10 Internet Movie Database

This dataset details the co-acting network of for movies released after December 31st 1999 as compiled by IMDb [88].

TABLE B.19 – IMDb statistics

N	L	$k_{\max}$	$c_{\max}$	$b_{\max}$	$m_{\max}$	ρ
716463	7665259	4625	192	N/A	2152	0.52

TABLE B.20 – IMDb correlations

$\sigma(k,m)$	$\sigma(b,m)$	$\sigma(c,m)$	$\sigma(k,b)$	$\sigma(k,c)$	$\sigma(b,c)$
0.6830	N/A	0.6186	N/A	0.9813	N/A

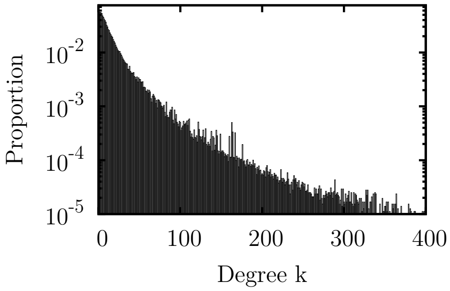


FIGURE B.27 – IMDb degree distribution

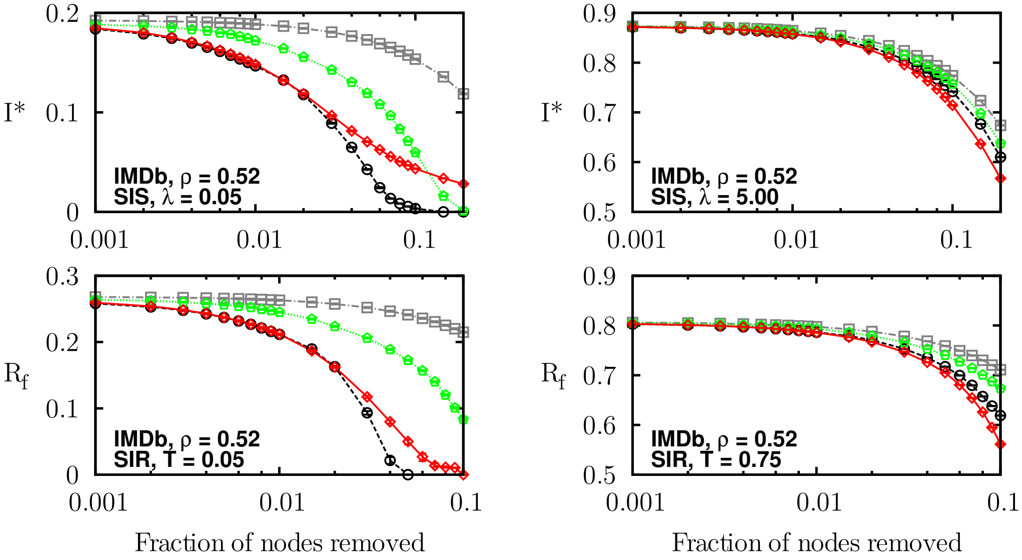


FIGURE B.28 – Intervention against epidemics on IMDb network after different immunization: randomly (grey squares) and based on coreness (green pentagons), degree (black circles), betweenness centrality (blue triangles) or memberships (red diamonds).

B.9.11 MathSciNet co-authorship

Co-authorship network of MathSciNet before 2008 [155].

TABLE B.21 – MathSci statistics

N	L	$k_{\max}$	$c_{\max}$	$b_{\max}$	$m_{\max}$	ρ
391529	873775	496	24	$1.9e+09$	485	0.40

TABLE B.22 – MathSci correlations

$\sigma(k,m)$	$\sigma(b,m)$	$\sigma(c,m)$	$\sigma(k,b)$	$\sigma(k,c)$	$\sigma(b,c)$
0.8645	0.8320	0.7749	0.7835	0.9465	0.6200

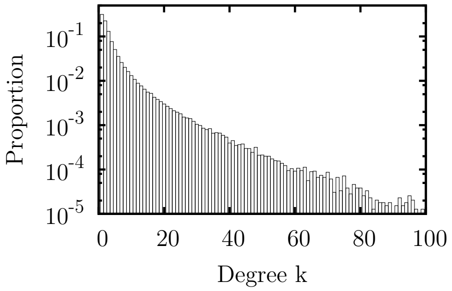


FIGURE B.29 – MathSci degree distribution

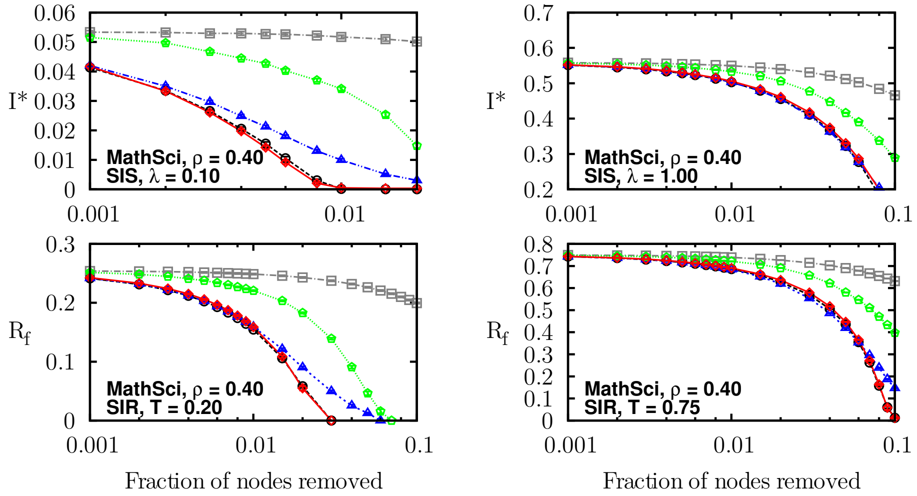


FIGURE B.30 – Intervention against epidemics on MathSci network after different immunization: randomly (grey squares) and based on coreness (green pentagons), degree (black circles), betweenness centrality (blue triangles) or memberships (red diamonds).

B.9.12 Myspace online social network

Friendships between the first 100,000 users encountered while crawling Myspace accounts from September to October 2006 (excluding Tom Anderson, the cofounder of MySpace, which is connected to everyone) [3].

TABLE B.23 – Myspace statistics

N	L	$k_{\max}$	$c_{\max}$	$b_{\max}$	$m_{\max}$	ρ
100000	841224	59108	78	$2.6e+09$	59102	0.77

TABLE B.24 – Myspace correlations

$\sigma(k,m)$	$\sigma(b,m)$	$\sigma(c,m)$	$\sigma(k,b)$	$\sigma(k,c)$	$\sigma(b,c)$
1.0000	0.8667	0.9995	0.8667	0.9995	0.8662

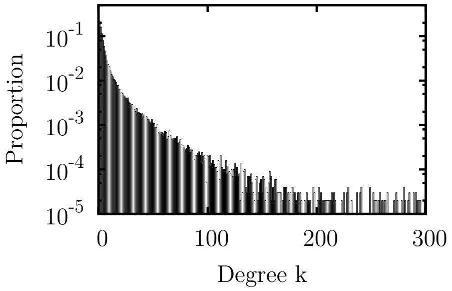


FIGURE B.31 – Myspace degree distribution

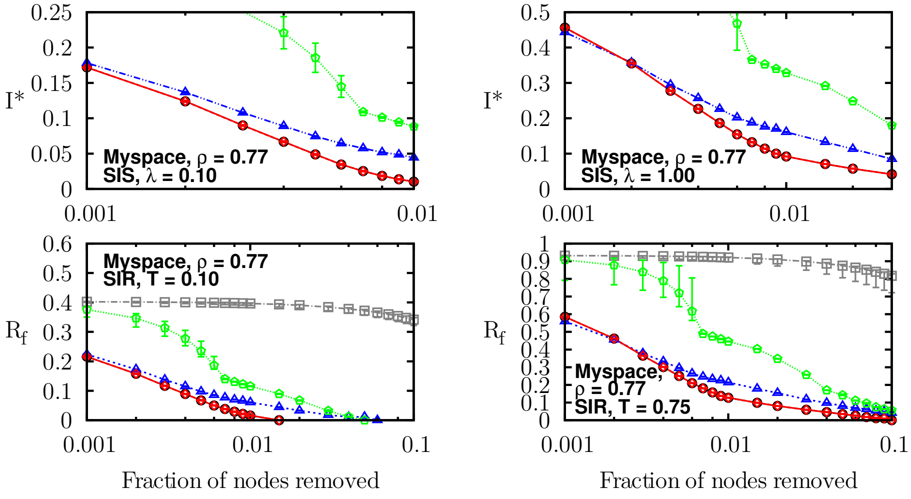


FIGURE B.32 – Intervention against epidemics on Myspace after different immunization: randomly (grey squares) and based on coreness (green pentagons), degree (black circles), betweenness centrality (blue triangles) or memberships (red diamonds).

B.9.13 Pretty-Good-Privacy data exchange

Dataset describing the giant component in the network of users of the Pretty-Good-Privacy algorithm for information exchange [26].

TABLE B.25 – PGP statistics

N	L	$k_{\max}$	$c_{\max}$	$b_{\max}$	$m_{\max}$	ρ
10680	24316	205	31	$7.5e+06$	110	0.50

TABLE B.26 – PGP correlations

$\sigma(k,m)$	$\sigma(b,m)$	$\sigma(c,m)$	$\sigma(k,b)$	$\sigma(k,c)$	$\sigma(b,c)$
0.8862	0.8599	0.7256	0.7973	0.8973	0.5464

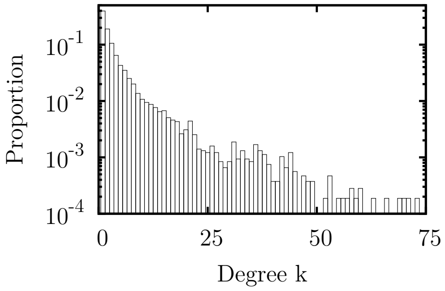


FIGURE B.33 – PGP degree distribution

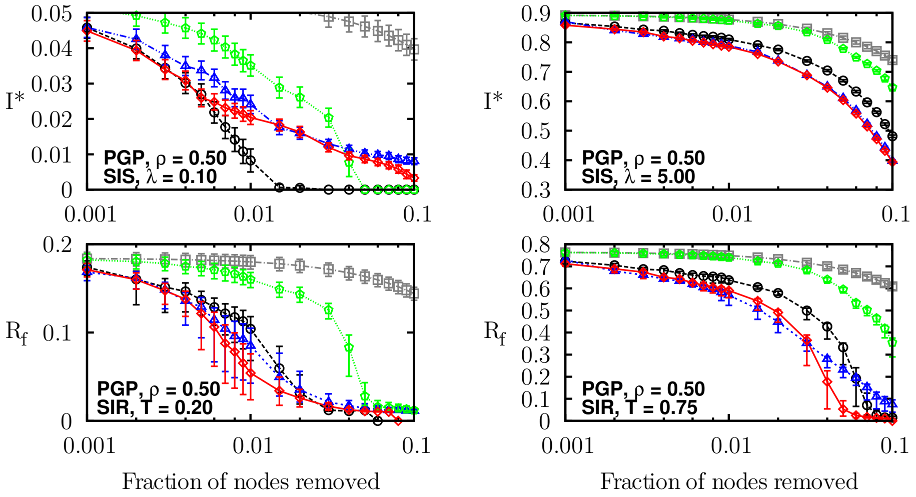


FIGURE B.34 – Intervention against epidemics on the PGP network after different immunization: randomly (grey squares) and based on coreness (green pentagons), degree (black circles), betweenness centrality (blue triangles) or memberships (red diamonds).

B.9.14 Power grid

The topology of the Western States Power Grid of the United States as compiled by Duncan Watts and Steven Strogatz [197].

TABLE B.27 – Power grid statistics

N	L	$k_{\max}$	$c_{\max}$	$b_{\max}$	$m_{\max}$	ρ
4941	6594	19	5	$3.5e+06$	18	0.49

TABLE B.28 – Power grid correlations

$\sigma(k,m)$	$\sigma(b,m)$	$\sigma(c,m)$	$\sigma(k,b)$	$\sigma(k,c)$	$\sigma(b,c)$
0.9192	0.8605	0.6191	0.8042	0.7342	0.5788

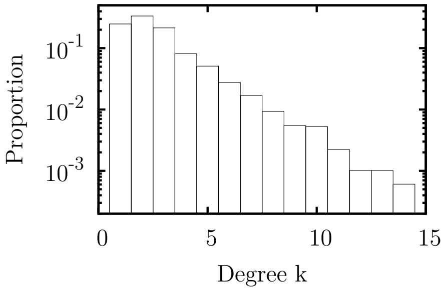


FIGURE B.35 – Power grid degree distribution

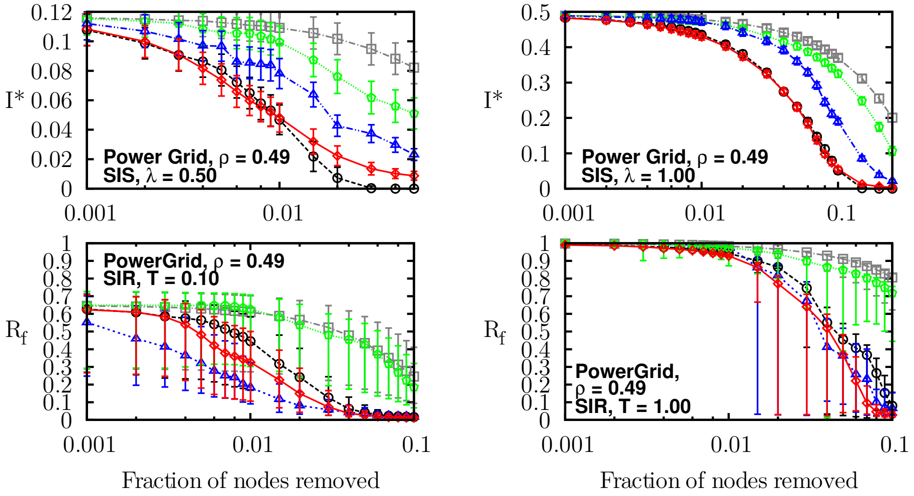


FIGURE B.36 – Intervention against epidemics on the power grid after different immunization: randomly (grey squares) and based on coreness (green pentagons), degree (black circles), betweenness centrality (blue triangles) or memberships (red diamonds).

B.9.15 Protein interactions network

Protein-protein interactions (ProteinCore) in *S. cerevisiae* as listed by the Database of Interacting Proteins [153].

TABLE B.29 – ProteinCore statistics

N	L	$k_{\max}$	$c_{\max}$	$b_{\max}$	$m_{\max}$	ρ
2640	6600	111	8	$4.0e+05$	71	0.32

TABLE B.30 – ProteinCore correlations

$\sigma(k,m)$	$\sigma(b,m)$	$\sigma(c,m)$	$\sigma(k,b)$	$\sigma(k,c)$	$\sigma(b,c)$
0.8538	0.9118	0.7702	0.8828	0.9543	0.7712

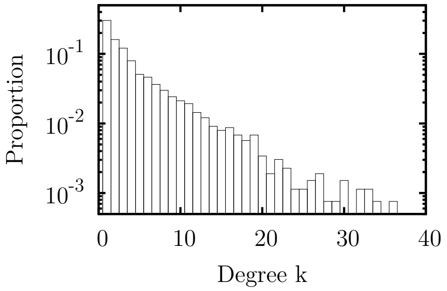


FIGURE B.37 – ProteinCore degree distribution

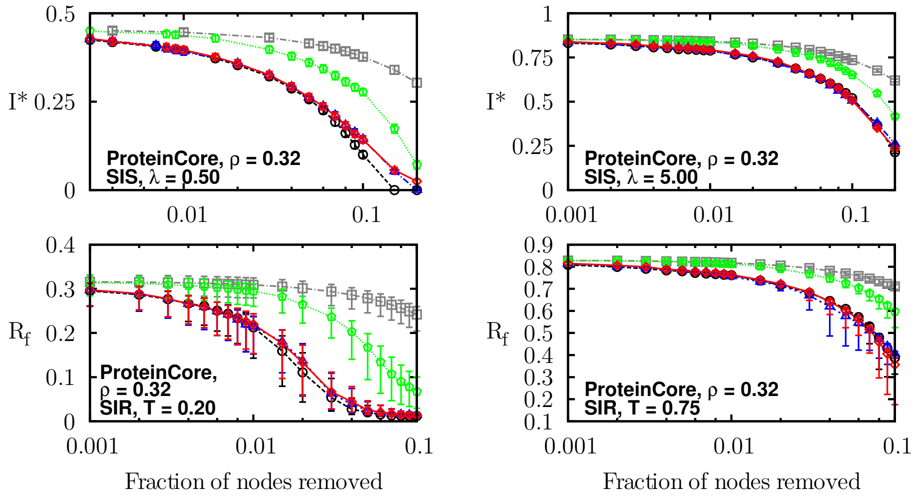


FIGURE B.38 – Intervention against epidemics on the protein interactions network after different immunization: randomly (grey squares) and based on coreness (green pentagons), degree (black circles), betweenness centrality (blue triangles) or memberships (red diamonds).

B.9.16 Slashdot online social network

Network of tagged relationships (friends or foes) in the community of the Slashdot news website in November 2008 [110].

TABLE B.31 – Slashdot statistics

N	L	$k_{\max}$	$c_{\max}$	$b_{\max}$	$m_{\max}$	ρ
77360	469180	2539	54	$1.2e+08$	2506	0.46

TABLE B.32 – Slashdot correlations

$\sigma(k,m)$	$\sigma(b,m)$	$\sigma(c,m)$	$\sigma(k,b)$	$\sigma(k,c)$	$\sigma(b,c)$
0.9958	0.9373	0.9832	0.9358	0.9870	0.8855

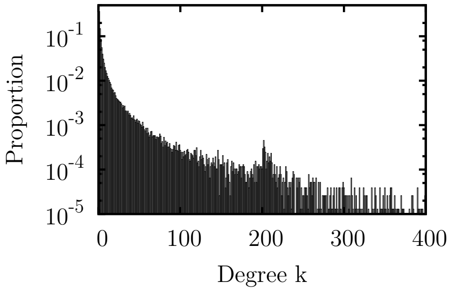


FIGURE B.39 – Slashdot degree distribution

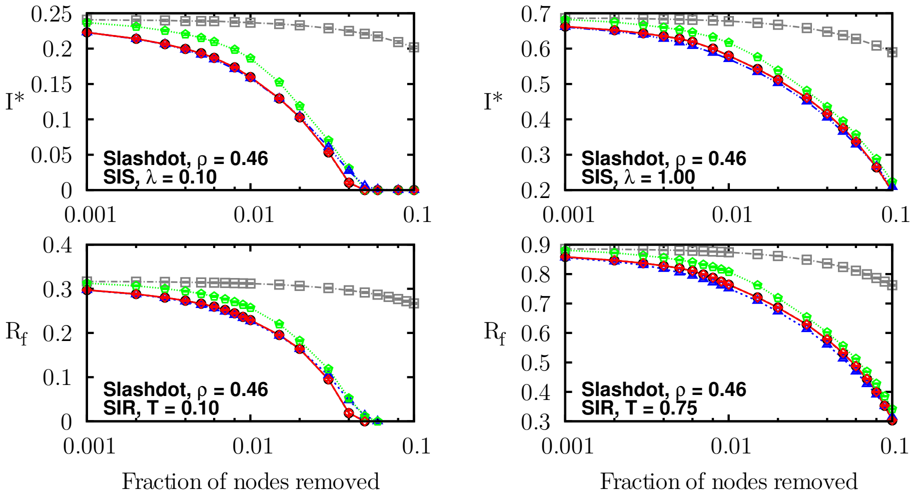


FIGURE B.40 – Intervention against epidemics on Slashdot after different immunization: randomly (grey squares) and based on coreness (green pentagons), degree (black circles), betweenness centrality (blue triangles) or memberships (red diamonds).

B.9.17 Word association network

Word association graph (built by survey) obtained from the South Florida Free Association norms [153].

TABLE B.33 – Word ass. statistics

N	L	$k_{\max}$	$c_{\max}$	$b_{\max}$	$m_{\max}$	ρ
7207	31784	218	7	$1.2e+06$	137	0.16

TABLE B.34 – Word ass. correlations

$\sigma(k,m)$	$\sigma(b,m)$	$\sigma(c,m)$	$\sigma(k,b)$	$\sigma(k,c)$	$\sigma(b,c)$
0.9698	0.9230	0.9110	0.9229	0.9281	0.8337

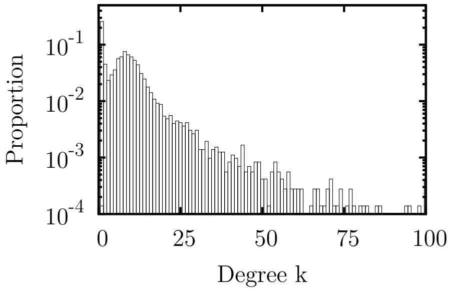


FIGURE B.41 – Word ass. degree distribution

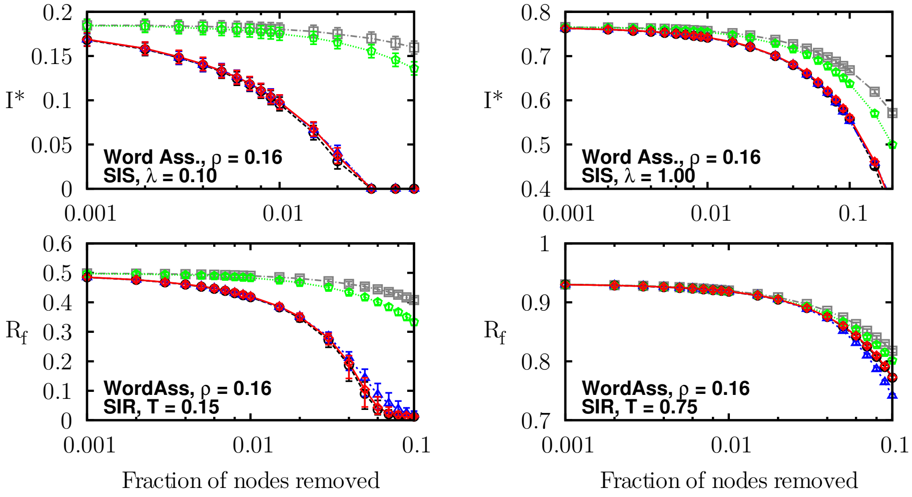


FIGURE B.42 – Intervention against epidemics on the word association graph after different immunization: randomly (grey squares) and based on coreness (green pentagons), degree (black circles), betweenness centrality (blue triangles) or memberships (red diamonds).

B.9.18 World Wide Web

Network of links between the webpages within *nd.edu* domain and considered undirected for this study [15].

TABLE B.35 – WWW statistics

N	L	$k_{\max}$	$c_{\max}$	$b_{\max}$	$m_{\max}$	ρ
325729	1090108	10721	155	$2.5e+10$	6993	0.86

TABLE B.36 – WWW correlations

$\sigma(k,m)$	$\sigma(b,m)$	$\sigma(c,m)$	$\sigma(k,b)$	$\sigma(k,c)$	$\sigma(b,c)$
0.9569	0.8683	0.9020	0.8665	0.9614	0.7905

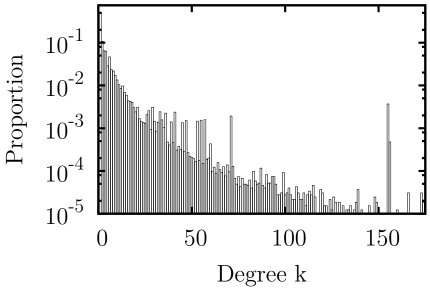


FIGURE B.43 – WWW degree distribution

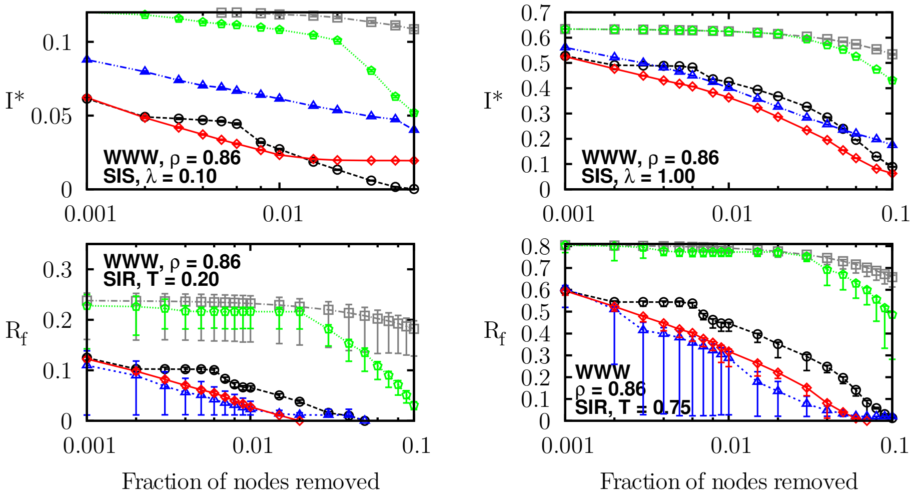


FIGURE B.44 – Intervention against epidemics on the WWW after different immunization: randomly (grey squares) and based on coreness (green pentagons), degree (black circles), betweenness centrality (blue triangles) or memberships (red diamonds).

Annexe C

Remarques sur la convergence des simulations numériques

Les résultats de simulations numériques faisant intervenir un processus stochastique quelconque (ex. de type *Monte Carlo*) prennent typiquement la forme d'une distribution. Cette distribution peut être inhérente au processus stochastique (ex. voir les figures 2.1b et 3.3), ou peut être due, par exemple, à un effet de taille finie du système modélisé (ex. la taille de la plus grande composante dans un graphe aléatoire suit une distribution qui s'apparente à une Gaussienne [149, 150]). L'analyse de la distribution en elle-même — ou de ses moments — offre généralement l'information nécessaire sur le processus stochastique modélisé [1]. Toutefois, lorsqu'il s'agit d'estimer la convergence de ladite distribution (c.-à-d. le nombre de réalisations du processus est-il suffisant ?), ses moments ne sont typiquement d'aucune utilité. La question est donc la suivante : comment peut-on juger de la convergence d'une distribution obtenue d'un processus stochastique ?

Nous nous restreignons à la situation dans laquelle un processus stochastique $\mathcal{P}$ correspond au système original à modéliser et où nous avons développé un modèle théorique $\mathcal{T}$ permettant de prédire la distribution résultante. Ce modèle peut être exact ou approximatif ; nous cherchons à le valider en comparant ses prédictions aux résultats numériques provenant du processus $\mathcal{P}$. L'objectif est de quantifier les fluctuations dans la distribution qui sont inhérentes à la nature stochastique du processus $\mathcal{P}$ afin d'identifier les différences entre les résultats numériques et les prédictions théoriques dues aux approximations, ou aux erreurs, de l'approche théorique $\mathcal{T}$.

[1] Par exemple, suivre la valeur moyenne d'une observable de même que son écart-type en fonction du temps permet de juger de l'étalement des trajectoires d'un processus stochastique dans l'espace des phases correspondant.

C.1 Processus stochastique et nomenclature

Considérons le processus stochastique $\mathcal{P}$ duquel M états finaux distincts $i \in \{1, \ldots, M\}$ peuvent résulter. La distribution des états finaux $\{\tilde{p}_i\}$ telle qu'obtenue après N réalisations du processus $\mathcal{P}$ est calculée selon

$$\tilde{p}_i = \frac{n_i}{N} \,, \tag{C.1}$$

où n_i est le nombre de réalisations ayant mené à l'état final i. Nous supposons que chacune de ces réalisations est indépendante. Le modèle théorique $\mathcal{T}$ prédit quant à lui que l'état final i se produit avec une probabilité p_i et nous notons $\{p_i\}$ la distribution finale associée.

C.2 Approche empirique

Supposons que nous ayons simulé le processus stochastique $\mathcal{P}$ un très grand nombre de fois et que nous ayons donc en notre possession une tout aussi longue séquence $\mathcal{S}$ d'états finaux. La fréquence $\tilde{p}_i$ de chaque état final i peut être calculée directement à l'aide de l'équation (C.1). Pour estimer les fluctuations statistiques (ou la convergence) de cette distribution, une première approche consiste à séparer la séquence $\mathcal{S}$ en un grand nombre G de segments de même longueur $\mathcal{S}_j$ tel que $\mathcal{S} = \bigcup_{j=1}^{G} \mathcal{S}_j$. Pour chaque segment $\mathcal{S}_j$, on calcule la fréquence $\tilde{p}_i^{(j)}$ de chaque état final i à l'aide d'une équation analogue à l'équation (C.1). On évalue finalement l'écart quadratique moyen Δ_i entre les $\tilde{p}_i^{(j)}$ et $\tilde{p}_i$ selon

$$\Delta_i = \left[\frac{1}{G} \sum_{j=1}^{G} \left(\tilde{p}_i^{(j)} - \tilde{p}_i \right)^2 \right]^{\frac{1}{2}} \,. \tag{C.2}$$

Cette mesure nous permet d'estimer à quel point la distribution tirée d'une séquence quelconque de longueur $|\mathcal{S}_j|$ devrait être différente de toute autre séquence de même longueur issue du même processus stochastique. Pour un nombre de segments G suffisamment grand (pour éviter des fluctuations $\{\Delta_i\}$ fortuitement petites), nous pourrons être d'autant plus confiants de la convergence de la distribution que les fluctuations $\{\Delta_i\}$ seront petites.

Quoique intuitive, cette méthode possède quelques désavantages. D'une part, elle nécessite un très grand nombre de simulations (c.-à-d. fardeau numérique) pour ultimement estimer des fluctuations qui correspondent à celles des sous-segments $\mathcal{S}_j$ et non à celles de la séquence complète $\mathcal{S}$. Puisque la distribution $\{\tilde{p}_i\}$ calculée à l'aide de la séquence complète $\mathcal{S}$ aura convergé davantage que celles calculées à l'aide d'un segment $\mathcal{S}_i^{(j)}$, les fluctuations $\{\Delta_i\}$ n'offrent qu'une borne maximale aux véritables fluctuations de $\{\tilde{p}_i\}$. D'autre part, cette méthode n'offre aucun pouvoir prédictif ; il est impossible d'estimer *a priori* le nombre de réalisations nécessaires de $\mathcal{P}$ pour obtenir la convergence souhaitée.

C.3 Approche théorique

Nous proposons d'aborder cette problématique selon une autre perspective qui capitalise plutôt sur les prédictions du modèle théorique $\mathcal{T}$. En effet, plutôt que de considérer d'abord les résultats provenant du processus $\mathcal{P}$, de tenter d'en isoler les fluctuations statistiques et ensuite de les confronter aux prédictions théoriques, cette approche inverse les rôles prenant comme point de départ le modèle théorique $\mathcal{T}$ et un processus stochastique associé $\mathcal{A}$.

Rappelons que nous connaissons la distribution $\{p_i\}$ grâce au modèle théorique $\mathcal{T}$. Supposons un processus stochastique associé $\mathcal{A}$ où l'événement i se produit avec une probabilité p_i indépendamment des autres événements. Tel que démontré à la section C.3.1, une analyse mathématique simple nous permet d'estimer les fluctuations entre la distribution théorique $\{p_i\}$ et la distribution obtenue à la suite de N réalisations du processus stochastique $\mathcal{A}$. Si le modèle théorique $\mathcal{T}$ est exact au sens où il reproduit exactement la distribution finale du processus stochastique original $\mathcal{P}$ (voir les figures 2.1b, 3.3, 5.2, A.2 et B.8) alors il n'y a statistiquement pas de différence[2] entre les processus $\mathcal{A}$ et $\mathcal{P}$. Dès lors l'équation (C.7) permet de d'estimer *a priori* le nombre N de réalisations de $\mathcal{P}$ nécessaires pour obtenir une convergence à la précision souhaitée. Cette approche se soustrait donc aux deux reproches énumérés précédemment à l'égard de l'approche dite empirique.

Dans le cas où le modèle théorique $\mathcal{T}$ est inexact (voir les figures 5.4, 5.5, A.3 et A.4), cette approche permet d'estimer les fluctuations auxquelles nous serions en droit de nous attendre si $\mathcal{A}$ avait été statistiquement similaire à $\mathcal{P}$. Autrement dit, si cette approche indique que N est tel que les fluctuations de la distribution obtenue via $\mathcal{A}$ seraient encore trop grandes, il est fort probable que les résultats obtenus à l'aide de $\mathcal{P}$ soient également entachés de fluctuations similaires. Cette approche permet donc d'estimer rapidement la convergence d'une distribution obtenue à l'aide d'un processus stochastique. Une fois cette convergence atteinte, il est alors possible d'isoler les véritables différences (autres que les fluctuations statistiques) entre les prédictions théoriques de $\mathcal{T}$ et les résultats provenant de $\mathcal{P}$, et par conséquent d'identifier des pistes potentielles pour améliorer le modèle théorique $\mathcal{T}$.

C.3.1 Traitement mathématique : un dé imparfait à M faces

L'analyse du processus stochastique $\mathcal{A}$ est identique à celle de N lancers d'un dé imparfait à M faces pour lequel la face i est obtenue avec une probabilité p_i. Nous utiliserons cette analogie et nous noterons $\mathcal{E}$ l'ensemble des résultats possibles (c.-à-d. les faces). Étant donnée l'indépendance des lancers, la probabilité d'obtenir la distribution $\{n_i\}$ à la suite de N

[2] Cette affirmation n'est valide que si l'indépendance entre les événements est respectées pour les processus stochastiques $\mathcal{A}$ et $\mathcal{P}$.

lancers est donnée simplement par la distribution multinomiale

$$\Pr\left[\{n_i\}\middle|N\right] = N!\prod_{i\in\mathcal{E}}\frac{p_i^{n_i}}{n_i!}\ , \tag{C.3}$$

où n_i est le nombre de fois que la face i fut obtenue. À la suite de N lancers, il y a $\binom{N+M-1}{M-1}$ distributions possibles et nous noterons $\mathcal{N}$ l'ensemble correspondant. Il est implicite que $\sum_{i\in\mathcal{E}} n_i = N$.

Analyse de la convergence

Après N lancers, le nombre moyen de fois que la face i aura été obtenue s'obtient par différentiation paramétrique [163]

$$\begin{aligned}\langle n_i\rangle &= \sum_{\{n_i\}\in\mathcal{N}} n_i \Pr\left[\{n_i\}\middle|N\right] \\ &= \sum_{\{n_i\}\in\mathcal{N}}\left(p_i\frac{\partial}{\partial p_i}\right)\Pr\left[\{n_i\}\middle|N\right] \\ &= \left(p_i\frac{\partial}{\partial p_i}\right)\sum_{\{n_i\}\in\mathcal{N}}\Pr\left[\{n_i\}\middle|N\right] \\ &= \left(p_i\frac{\partial}{\partial p_i}\right)\left(\sum_{i\in\mathcal{E}}p_i\right)^N \\ &= Np_i\ ,\end{aligned} \tag{C.4}$$

où le théorème multinomial a été utilisé notamment pour inverser l'ordre de la somme et de l'opérateur différentiel (c.-à-d. que la somme converge) [67] et où nous avons utilisé le fait que $\sum_{i\in\mathcal{E}} p_i = 1$. Nous constatons sans grande surprise que la fréquence de la face i sera en moyenne égale à $\langle n_i\rangle/N = p_i$ après N lancers.

Ceci étant dit, la distribution $\{n_i/N\}$ obtenue après une seule série de N lancers sera différente de $\{p_i\}$, et cette différence s'estime à l'aide de l'écart-type. Pour ce faire, calculons $\langle n_i^2\rangle$ de façon similaire

$$\begin{aligned}\langle n_i^2\rangle &= \sum_{\{n_i\}\in\mathcal{N}} n_i^2 \Pr\left[\{n_i\}\middle|N\right] \\ &= \sum_{\{n_i\}\in\mathcal{N}}\left(p_i\frac{\partial}{\partial p_i}\right)^2\Pr\left[\{n_i\}\middle|N\right] \\ &= Np_i\left[1+(N-1)p_i\right]\ .\end{aligned} \tag{C.5}$$

En combinant les équations (C.4) et (C.5), nous obtenons l'écart-type moyen entre le nombre d'occurrences de la face i après N lancers et sa valeur théorique Np_i

$$\begin{aligned}\sigma_i &= \left[\left\langle n_i^2\right\rangle - \left\langle n_i\right\rangle^2\right]^{\frac{1}{2}} \\ &= \left[Np_i(1-p_i)\right]^{\frac{1}{2}} \ . \end{aligned} \tag{C.6}$$

Ainsi, nous voyons que l'écart relatif moyen, ε_i, entre n_i et sa valeur théorique Np_i se comportera selon

$$\varepsilon_i \equiv \frac{\left|\left(\left\langle n_i\right\rangle \pm \sigma_i\right) - \left\langle n_i\right\rangle\right|}{\left\langle n_i\right\rangle} = \sqrt{\frac{1-p_i}{p_i}}\frac{1}{\sqrt{N}} \tag{C.7}$$

$$\simeq \frac{1}{\sqrt{Np_i}} \qquad \text{si } p_i \ll 1 \ . \tag{C.8}$$

Ce résultat permet donc d'estimer le nombre de lancers N nécessaires pour que la distribution obtenue à l'aide du processus $\mathcal{A}$ ait convergé à une précision souhaitée.

Résultats agglomérés

Il est possible que la comparaison du processus $\mathcal{P}$ et des prédictions du modèle $\mathcal{T}$ ne se fasse pas directement via les distributions $\{\tilde{p}_i\}$ et $\{p_i\}$, mais plutôt via des distributions agglomérées (ex. la projection d'une distribution de probabilité jointe sur une de ses variables). Nous montrons dans cette section que l'analyse présentée dans la section précédente s'adapte directement à ce cas.

Considérons un sous-ensemble $\mathcal{M} \in \mathcal{E}$ contenant les éléments de la distribution $\{p_i\}$ que nous souhaitons agglomérer (dans ce cas-ci sommer). La valeur moyenne de cette somme s'obtient directement

$$\langle n_{\mathcal{M}}\rangle \equiv \left\langle \sum_{i\in\mathcal{M}} n_i \right\rangle = \sum_{i\in\mathcal{M}} \left\langle n_i\right\rangle = N\sum_{i\in\mathcal{M}} p_i \tag{C.9}$$

où nous avons utilisé (C.4) de même que la linéarité de l'opération $\left\langle \cdots \right\rangle$. Par une démarche similaire, nous obtenons

$$\begin{aligned}\left\langle \left(\sum_{i\in\mathcal{M}} n_i\right)^2 \right\rangle &= \sum_{i\in\mathcal{M}} \left\langle n_i^2\right\rangle + \sum_{\substack{j,k\in\mathcal{M} \\ j\neq k}} \left\langle n_j n_k\right\rangle \\ &= \sum_{i\in\mathcal{M}} \left\langle n_i^2\right\rangle + \sum_{\substack{j,k\in\mathcal{M} \\ j\neq k}} \left(\sum_{\{n_i\}\in\mathcal{N}} n_j n_k \mathrm{Pr}\left[\{n_i\}\middle|N\right]\right) \\ &= \sum_{i\in\mathcal{M}} Np_i\left[1+(N-1)p_i\right] + \sum_{\substack{j,k\in\mathcal{M} \\ j\neq k}} N(N-1)p_jp_k \\ &= N\left(\sum_{i\in\mathcal{M}} p_i\right) + N(N-1)\left(\sum_{i\in\mathcal{M}} p_i\right)^2 \ . \end{aligned} \tag{C.10}$$

Le passage de la seconde à la troisième égalité s'est fait en vertu d'une différentiation paramétrique sous la condition stricte que $j \neq k$. En combinant les résultats (C.9) et (C.10), nous obtenons l'écart quadratique moyen entre les données agglomérées $\mathcal{M}$ et leur valeur théorique donnée par (C.9)

$$\begin{aligned} \sigma_{\mathcal{M}} &= \left[\left\langle \left(\sum_{i \in \mathcal{M}} n_i \right)^2 \right\rangle - \left\langle \left(\sum_{i \in \mathcal{M}} n_i \right) \right\rangle^2 \right]^{\frac{1}{2}} \\ &= \left[N \left(\sum_{i \in \mathcal{M}} p_i \right) \left(1 - \sum_{i \in \mathcal{M}} p_i \right) \right]^{\frac{1}{2}} \\ &= \left[N p_{\mathcal{M}} (1 - p_{\mathcal{M}}) \right]^{\frac{1}{2}} \ , \end{aligned} \tag{C.11}$$

où nous avons noté $p_{\mathcal{M}} = \sum_{i \in \mathcal{M}} p_i$. Ce résultat est analogue à l'équation (C.6) ; la probabilité d'obtenir la face i dans (C.6) étant simplement remplacée par la somme des probabilités individuelles des données agglomérées de $\mathcal{M}$. Ainsi, nous voyons qu'il est possible d'analyser la convergence de distributions agglomérées avec la même approche que celle développée pour les données individuelles [3].

[3] Ceci s'explique par le fait que la somme des éléments de $\mathcal{M}$ est elle-même distribuée selon une distribution binomiale. En effet, on peut montrer que

$$\begin{aligned} \Pr\left[n_{\mathcal{M}} \middle| N \right] &= \sum_{\{n_i\} \in \mathcal{N}} \delta \left(n_{\mathcal{M}} - \sum_{i \in \mathcal{M}} n_i \right) \Pr\left[\{n_i\} \middle| N \right] \\ &= \binom{N}{n_{\mathcal{M}}} p_{\mathcal{M}}^{n_{\mathcal{M}}} (1 - p_{\mathcal{M}})^{N - n_{\mathcal{M}}} \ . \end{aligned}$$

Le résultat (C.11) découle alors directement.

Annexe D

Remarques sur la solution numérique d'équations

En général, les équations permettant de prédire les propriétés structurelles (ex. taille de la composante géante, seuil de percolation) des graphes présentés aux chapitres 1–5 et aux annexes A–B ne possèdent pas de solutions analytiques explicites et nécessitent, par conséquent, l'utilisation de méthodes numériques pour être résolues. Nous décrivons brièvement les méthodes utilisées pour obtenir les résultats présentés dans cette thèse.

Graphes arbitraires de petite taille—Les distributions de la taille des composantes sur des graphes de petite taille s'obtiennent directement en itérant une condition initiale appropriée à l'aide des équations de récurrence (1.28), (1.29), (3.17), (3.18) et (4.1).

Seuil de percolation—Lorsque les seuils de percolation furent calculés explicitement en terme d'un paramètre donné (ex. voir les figures 3.2 et 5.2), la valeur critique de ce paramètre fut obtenue en résolvant les équations (3.7), (4.9), (5.14), (5.17) et (5.26) à l'aide de la méthode de bissection [68].

Composante géante—Les quantités relatives à la composante géante (ex. taille relative, composition et probabilité d'existence) s'obtiennent en trouvant le point fixe stable des équations (1.24), (1.64), (3.8), (3.10), (4.7), (4.11), (4.16), (5.12), (5.15), (5.23), (5.25), (A.10) et (B.12). Bien qu'il existe des techniques plus rapides et efficaces pour obtenir la position des points fixes de ce type d'équations, nous les avons obtenus en itérant une condition initiale quelconque dans l'intervalle $[0, 1]$ jusqu'à ce que les itérés aient convergé à la précision souhaitée (typiquement de l'ordre de 10^{-8}).

Taille des petites composantes—La distribution de la taille des petites composantes s'obtient en résolvant les équations (1.78), (3.13) et (4.13). Ces équations peuvent être résolues en considérant les systèmes dynamiques associés [ex. les équations (1.79) et (4.14)] avec lesquels on itère une condition initiale appropriée jusqu'à ce que la distribution soit obtenue jusqu'à une taille maximale de composante désirée (ex. la distribution est obtenue pour les composantes contenant 25 nœuds ou moins). Cette façon de faire a été déjà été implémentée dans un autre contexte (voir l'annexe B de [5] pour les détails), et bien qu'en principe cette méthode soit valide en tout temps, elle devient rapidement lourde au point de vue numérique dès qu'il y a plus de deux types de nœuds (c.-à-d. mémoire et temps de calcul requis).

Pour obtenir les résultats analytiques montrés à la figure 3.3, nous avons eu recours à une autre méthode qui permet de contourner cette lourdeur numérique [38, 118, 132, 147, 148]. Nous illustrons cette méthode à l'aide des équations (1.75) et (1.81) que nous réécrivons ici

$$A(z) = zf\big(A(z)\big) \tag{D.1}$$

$$K(z) = zg\big(A(z)\big)\ . \tag{D.2}$$

Tel qu'expliqué à la section 1.3.2, la distribution de la taille des petites composantes, $\{\eta(s)\}$, s'obtient par différentiation de la fonction génératrice associée

$$\eta(s) = \left[\frac{1}{s!}\frac{\partial^s}{\partial z^s}K(z)\right]_{z=0}\ . \tag{D.3}$$

Du théorème intégral de Cauchy, nous pouvons récrire ces dérivées en terme d'une intégrale de parcours dans le plan complexe

$$\eta(s) = \frac{1}{2i\pi}\oint_\Gamma \frac{K(z)}{z^{s+1}}dz\ , \tag{D.4}$$

où Γ est un parcours circulaire de rayon r centré à l'origine et circonscrivant un disque dans lequel la fonction $K(z)$ est analytique. Ainsi, en effectuant la substitution $z = re^{2i\pi\phi}$ et en discrétisant l'intégrale en N points également espacés, l'équation (D.4) devient

$$\eta(s) = \frac{1}{2i\pi}\oint_\Gamma \frac{K(z)}{z^{s+1}}dz = \frac{1}{r^s}\int_0^1 K\big(re^{2i\pi\phi}\big)\mathrm{e}^{-2i\pi s\phi}d\phi \simeq \frac{1}{Nr^s}\sum_{n=0}^{N-1} K\big(re^{2i\pi n/N}\big)\mathrm{e}^{-2i\pi sn/N}\ , \tag{D.5}$$

où nous identifions une transformée de Fourier inverse discrète de la séquence $\{K\big(re^{2i\pi n/N}\big)\}$ avec $n = 0, \ldots, N-1$ [161]. Bien que la forme exacte de la fonction génératrice $K(z)$ ne soit pas connue (cet exercice serait alors inutile), il est tout de même possible de connaître sa valeur aux N points équidistants sur le cercle unitaire à l'aide des équations (D.1) et (D.2). En effet, il suffit de calculer la valeur des $A_n \equiv A(re^{2i\pi n/N})$ satisfaisant l'équation (D.1) à l'aide de l'algorithme de notre choix permettant de trouver les zéros d'une fonction, et ensuite d'injecter ceux-ci dans l'équation (D.2). La probabilité $\eta(s)$ correspond en quelque sorte à l'amplitude de la fréquence s de la séquence $\{K\big(re^{2i\pi n/N}\big)\}$. Il est donc possible de calculer l'ensemble de ces coefficients en une seule passe d'un algorithme FFT [161] : une amélioration considérable du temps de calcul par rapport à la façon de faire précédente (c.-à-d. au calcul par itération explicite). Notons également que cette technique se généralise directement à plusieurs dimensions : l'équation (D.5) correspond alors à une transformée de Fourier inverse discrète multidimensionnelle.

Une attention particulière doit être portée au nombre de points, N, utilisés pour discrétiser l'intégrale dans l'équation (D.4). En effet, en vertu du théorème d'échantillonnage de Nyquist-Shannon, si $N > s_{\max}$, où $s_{\max}$ est le degré de $K(z)$ tel que $\eta(s) = 0$ pour tout $s > s_{\max}$, alors la distribution ainsi obtenue est en principe exacte. La précision numérique des $\eta(s)$ est alors déterminée par la précision avec laquelle les A_n ont été calculés et par le choix du parcours d'intégration (le cercle unitaire, c.-à-d. $r = 1$, permet en général d'obtenir une précision suffisante) [147]. Lorsque $N \leq s_{\max}$, les termes $\eta(s)$ pour lesquels $s \geq N$ sont « rabattus » sur les termes $\eta(s)$ pour lesquels $s < N$, induisant ainsi une erreur dans les valeurs calculées (phénomène d'*aliasing* [161]). Le degré de la fonction $K(z)$ est formellement infini, de sorte que nous sommes toujours en présence d'*aliasing*. Toutefois, si la distribution sous-tendue par $K(z)$ tombe suffisamment rapidement, il est possible de choisir une valeur de N telle que l'erreur induite par l'*aliasing* n'affecte pas significativement la précision du calcul.

Bibliographie

[1] M. ABRAMOWITZ et I. A. STEGUN : *Handbook of Mathematical Functions*. Dover, 9$^{\mathrm{e}}$ édition, 1972.

[2] Y.-Y. AHN, J. P. BAGROW et S. LEHMANN : Link communities reveal multiscale complexity in networks. *Nature*, 466:761–4, 2010.

[3] Y.-Y. AHN, S. HAN, H. KWAK, S. MOON et H. JEONG : Analysis of topological characteristics of huge online social networking services. *Dans Proc. 16th Int. Conf. World Wide Web*, pages 835–844, New York, 2007. ACM Press.

[4] R. ALBERT, H. JEONG et A.-L. BARABÁSI : Error and attack tolerance of complex networks. *Nature*, 406(6794):378–82, 2000.

[5] A. ALLARD : *Modélisation Mathématique en Epidémiologie par Réseaux de Contacts : Introduction de l'Hétérogénéité dans la Transmissibilité*. Mémoire de maîtrise, Université Laval, 2008.

[6] A. ALLARD, L. HÉBERT-DUFRESNE, P.-A. NOËL, V. MARCEAU et L. J. DUBÉ : Bond percolation on a class of correlated and clustered random graphs. *J. Phys. A*, 45(40): 405005, 2012.

[7] A. ALLARD, L. HÉBERT-DUFRESNE, P.-A. NOËL, V. MARCEAU et L. J. DUBÉ : Exact solution of bond percolation on small arbitrary graphs. *EPL*, 98(1):16001, 2012.

[8] A. ALLARD, L. HÉBERT-DUFRESNE, J.-G. YOUNG et L. J. DUBÉ : A general and exact approach to percolation on random graphs. *En préparation*

[9] A. ALLARD, L. HÉBERT-DUFRESNE, J.-G. YOUNG et L. J. DUBÉ : Coexistence of phases and the observability of random graphs. *Phys. Rev. E*, 89(2):022801, 2014.

[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(3):036113, 2009.

[11] J. I. ALVAREZ-HAMELIN, L. DALL'ASTA, A. BARRAT et A. VESPIGNANI : Large scale networks fingerprinting and visualization using the k-core decomposition. *Dans*

Y. WEISS, B. SCHÖLKOPF et J. PLATT, éditeurs : *Adv. Neural Inf. Process. Syst. 18*, pages 41–50, Cambridge, MA, 2005. MIT Press.

[12] R. M. ANDERSON et R. M. MAY : *Infectious Diseases of Humans : Dynamics and Control*. Oxford University Press, Oxford, 1991.

[13] A. ARENAS, A. DÍAZ-GUILERA, J. KURTHS, Y. MORENO et Z. CHANGSONG : Synchronization in complex networks. *Phys. Rep.*, 469(3):93–153, 2008.

[14] S. BANSAL, B. POURBOHLOUL et L. A. MEYERS : A comparative analysis of influenza vaccination programs. *PLoS Med.*, 3(10):e387, 2006.

[15] A.-L. BARABÁSI et R. ALBERT : Emergence of Scaling in Random Networks. *Science*, 286:509–512, 1999.

[16] M. BARTHÉLEMY : Betweenness centrality in large complex networks. *Eur. Phys. J. B*, 38(2):163–168, 2004.

[17] M. BARTHÉLEMY : Spatial networks. *Phys. Rep.*, 499:1–101, 2011.

[18] A. BASHAN, Y. BEREZIN, S. V. BULDYREV et S. HAVLIN : The extreme vulnerability of interdependent spatially embedded networks. *Nat. Phys.*, 9(10):667–672, 2013.

[19] V. BATAGELJ et M. ZAVERŠNIK : Generalized Cores. *arXiv :cs/0202039*, 2002.

[20] V. BATAGELJ et M. ZAVERŠNIK : An O(m) Algorithm for Cores Decomposition of Networks. *arXiv :cs.DS/0310049*, 2003.

[21] P. S. BEARMAN, J. MOODY et K. STOVEL : Chains of affection : The structure of adolescent romantic and sexual networks. *Am. J. Sociol.*, 110(1):44–91, 2004.

[22] E. A. BENDER et E. R. CANFIELD : The asymptotic number of labeled graphs degree sequences. *J. Comb. Theory Ser. A*, 24:296–307, 1978.

[23] Y. BERCHENKO, Y. ARTZY-RANDRUP, M. TEICHER et L. STONE : Emergence and size of the giant component in clustered random graphs with a given degree distribution. *Phys. Rev. Lett.*, 102(13):138701, 2009.

[24] H. A. BETHE : Statistical theory of superlattices. *Proc. R. Soc. London. Ser. A*, 150(871): 552–575, 1935.

[25] G. BIANCONI : Statistical mechanics of multiplex networks : Entropy and overlap. *Phys. Rev. E*, 87(6):062806, 2013.

[26] M. BOGUÑÁ, R. PASTOR-SATORRAS, A. DÍAZ-GUILERA et A. ARENAS : Models of social networks based on social distance attachment. *Phys. Rev. E*, 70(5):056122, 2004.

[27] M. BOGUÑÁ, R. PASTOR-SATORRAS et A. VESPIGNANI : Absence of epidemic threshold in scale-free networks with degree correlations. *Phys. Rev. Lett.*, 90(2):028701, 2003.

[28] B. BOLLOBÁS : A probabilistic proof of an asymptotic formula for the number of labelled regular graphs. *Eur. J. Comb.*, 1(4):311–316, 1980.

[29] B. BOLLOBÁS : *Random Graphs*. Cambridge University Press, second édition, 2001.

[30] U. BRANDES : A faster algorithm for betweenness centrality. *J. Math. Sociol.*, 25(2):163–177, 2001.

[31] S. R. BROADBENT et J. M. HAMMERSLEY : Percolation processes. *Math. Proc. Cambridge Philos. Soc.*, 53(3):629, 1957.

[32] Andrei BRODER, R. KUMAR, F. MAGHOUL, P. RAGHAVAN, S. RAJAGOPALAN, R. STATA, A. TOMKINS et J. WIENER : Graph structure in the Web. *Comput. Networks*, 33:309–320, 2000.

[33] C. D. BRUMMITT, R. M. D'SOUZA et E. A. LEICHT : Suppressing cascades of load in interdependent networks. *Proc. Natl. Acad. Sci. U. S. A.*, 109(12):E680–9, 2012.

[34] S. V. BULDYREV, R. PARSHANI, G. PAUL, H. E. STANLEY et S. HAVLIN : Catastrophic cascade of failures in interdependent networks. *Nature*, 464(7291):1025–8, 2010.

[35] G. CALDARELLI et A. VESPIGNANI : *Large Scale Structure and Dynamics of Complex Networks*. World Scientific, Singapore, 2007.

[36] D. S. CALLAWAY, J. E. HOPCROFT, J. M. KLEINBERG, M. E. J. NEWMAN et S. H. STROGATZ : Are randomly grown graphs really random ? *Phys. Rev. E*, 64(4):041902, 2001.

[37] 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(25):5468–71, 2000.

[38] J. K. CAVERS : On the Fast Fourier Transform Inversion of Probability Generating Functions. *J. Inst. Maths Applics*, 22:275–282, 1978.

[39] Y. CHEN, G. PAUL, S. HAVLIN, F. LILJEROS et H. EUGENE STANLEY : Finding a better immunization strategy. *Phys. Rev. Lett.*, 101(5):058701, 2008.

[40] A. CHO : Network science at center of surveillance dispute. *Science*, 340:1272, 2013.

[41] E. CHO, S. A. MYERS et J. LESKOVEC : Friendship and mobility : User movement in location-based social networks. *Dans Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pages 1082–1090, 2011.

[42] K. CHRISTENSEN et N. R. MOLONEY : *Complexity and Criticality*. Imperial College Press, London, 2005.

[43] A. CLAUSET, C. R. SHALIZI et M. E. J. NEWMAN : Power-law distributions in empirical data. *SIAM Rev.*, 51(4):661, 2009.

[44] R. COHEN et S. HAVLIN : *Complex Networks : Structure, Robustness and Function*. Cambridge University Press, 2010.

[45] P. COLOMER-DE-SIMÓN et M. BOGUÑÁ : Double percolation phase transition in clustered complex networks. *arXiv :1401.8176*, 2014.

[46] P. COLOMER-DE-SIMÓN, M. Á. SERRANO, M. G. BEIRÓ, J. I. ALVAREZ-HAMELIN et M. BOGUÑÁ : Deciphering the global organization of clustering in real complex networks. *Sci. Rep.*, 3:2517, 2013.

[47] N. J. COWAN, E. J. CHASTAIN, D. A. VILHENA, J. S. FREUDENBERG et C. T. BERGSTROM : Nodal dynamics, not degree distributions, determine the structural controllability of complex networks. *PLoS One*, 7(6):e38398, 2012.

[48] E. COZZO, M. KIVELÄ, M. DE DOMENICO, A. SOLÉ, A. ARENAS, S. GÓMEZ, M. A. PORTER et Y. MORENO : Clustering Coefficients in Multiplex Networks. *arXiv :1307.6780*, 2013.

[49] A. DÁVILA, C. R. ESCUDERO, J. A. LÓPEZ et C. O. DORSO : Geometrical aspects of isoscaling. *Physica A*, 374(2):663–668, 2007.

[50] S. DAVIS, P. TRAPMAN, H. LEIRS, M. BEGON et J. A. P. HEESTERBEEK : The abundance threshold for plague as a critical percolation phenomenon. *Nature*, 454(7204):634–7, 2008.

[51] M. DE CHOUDHURY, H. SUNDARAM, A. JOHN et D. D. SELIGMANN : Social synchrony : Predicting mimicry of user actions in online social media. *In 2009 Int. Conf. Comput. Sci. Eng.*, pages 151–158. IEEE, 2009.

[52] M. DE DOMENICO, A. SOLÉ-RIBALTA, E. COZZO, M. KIVELÄ, Y. MORENO, M. PORTER, S. GÓMEZ et A. ARENAS : Mathematical Formulation of Multilayer Networks. *Phys. Rev. X*, 3(4):041022, 2013.

[53] P. DÉSESQUELLES : Exact solution of finite size mean field percolation and application to nuclear fragmentation. *Phys. Lett. B*, 698(4):284–287, 2011.

[54] P. DÉSESQUELLES, D.-T. NGA et A. CHOLET : Exact solution of random graph fragmentation and physical, chemical and biological applications. *J. Phys. Conf. Ser.*, 410:012058, 2013.

[55] R. DIESTEL : *Graph Theory*. Springer-Verlag, Heidelberg, 4^{e} édition, 2012.

[56] S. N. DOROGOVTSEV : *Lectures on Complex Networks*. Oxford University Press, 2010.

[57] S. N. DOROGOVTSEV, A. V. GOLTSEV et J. F. F. MENDES : k-core organization of complex networks. *Phys. Rev. Lett.*, 96(4):040601, 2006.

[58] S. N. DOROGOVTSEV, A. V. GOLTSEV et J. F. F. MENDES : Critical phenomena in complex networks. *Rev. Mod. Phys.*, 80(4):1275–1335, 2008.

[59] S. N. DOROGOVTSEV et J. F. F. MENDES : *Evolution of Networks : From Biological Nets to the Internet and WWW*. Oxford University Press, 2003.

[60] J. A. DUNNE et R. J. WILLIAMS : Cascading extinctions and community collapse in model food webs. *Phil. Trans. R. Soc. B*, 364(1524):1711–23, 2009.

[61] M. ERCSEY-RAVASZ, R. N. LICHTENWALTER, N. V. CHAWLA et Z. TOROCZKAI : Range-limited centrality measures in complex networks. *Phys. Rev. E*, 85(6):066103, 2012.

[62] P. ERDŐS et A. RÉNYI : On random graphs I. *Publ. Math. Debrecen*, 6:290–297, 1959.

[63] P. ERDŐS et A. RÉNYI : On the evolution of random graphs. *Magy. Tud. Akad. Mat. Kut. Int. Közl.*, 5:17–61, 1960.

[64] P. ERDŐS et A. RÉNYI : On the evolution of random graphs. *Bull. Inst. Internat. Stat.*, 38:343–347, 1961.

[65] P. ERDŐS et A. RÉNYI : On the strength of connectedness of a random graph. *Acta Math. Acad. Sci. Hungar.*, 12:261–7, 1961.

[66] J. W. ESSAM : Percolation theory. *Rep. Prog. Phys.*, 43:833–912, 1980.

[67] W. FELLER : *An Introduction to Probability Theory and its Applications*. Wiley, 1968.

[68] A. FORTIN : *Analyse numérique pour ingénieurs*. Presses internationales Polytechnique, Montréal, Québec, Canada, deuxième édition, 2001.

[69] S. FORTUNATO : Community detection in graphs. *Phys. Rep.*, 486:75–174, 2010.

[70] G. F. FRASCO, J. SUN, H. D. ROZENFELD et D. BEN-AVRAHAM : Spatially distributed social complex networks. *Phys. Rev. X*, 4(1):011008, 2013.

[71] Linton C FREEMAN : Centrality in social networks conceptual clarification. *Soc. Networks*, 1:215–239, 1979.

[72] S. FUNK et V. A. A. JANSEN : Interacting epidemics on overlay networks. *Phys. Rev. E*, 81(3):036118, 2010.

[73] F. GAITAN et L. CLARK : Ramsey numbers and adiabatic quantum computing. *Phys. Rev. Lett.*, 108(1):010501, 2012.

[74] L. K. GALLOS, R. COHEN, P. ARGYRAKIS, A. BUNDE et S. HAVLIN : Stability and topology of scale-free networks under attack and defense strategies. *Phys. Rev. Lett.*, 94(18):188701, 2005.

[75] L. K. GALLOS, F. LILJEROS, P. ARGYRAKIS, A. BUNDE et S. HAVLIN : Improving immunization strategies. *Phys. Rev. E*, 75(4):045104(R), 2007.

[76] J. GAO, S. V. BULDYREV, H. E. STANLEY et S. HAVLIN : Networks formed from interdependent networks. *Nat. Phys.*, 8(1):40–48, 2011.

[77] G. GHOSHAL, V. ZLATIĆ, G. CALDARELLI et M. E. J. NEWMAN : Random hypergraphs and their applications. *Phys. Rev. E*, 79(6):066118, 2009.

[78] E. N. GILBERT : Random graphs. *Ann. Math. Stat.*, 30:1141–4, 1959.

[79] M. GIRVAN et M. E. J. NEWMAN : Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA*, 99(12):7821–6, 2002.

[80] J. P. GLEESON : Bond percolation on a class of clustered random networks. *Phys. Rev. E*, 80(3):036107, 2009.

[81] J. P. GLEESON, S. MELNIK et A. HACKETT : How clustering affects the bond percolation threshold in complex networks. *Phys. Rev. E*, 81(6):066114, 2010.

[82] J. GÓMEZ-GARDEÑES, P. ECHENIQUE et Y. MORENO : Immunization of real complex communication networks. *Eur. Phys. J. B*, 49(2):259–264, 2006.

[83] R. GUIMERÀ, L. DANON, A. DÍAZ-GUILERA, F. GIRALT et A. ARENAS : Self-similar community structure in a network of human interactions. *Phys. Rev. E*, 68(6):065103(R), 2003.

[84] A. HACKETT, J. P. GLEESON et S. MELNIK : Site percolation in clustered random networks. *Int. J. Comp. Syst. Sci.*, 1:25–32, 2011.

[85] J. HARRIS, J. L. HIRST et M. MOSSINGHOFF : *Combinatorics and Graph Theory*. Springer, 2^{e} édition, 2008.

[86] T. HASEGAWA, T. TAKAGUCHI et N. MASUDA : Observability transitions in correlated networks. *Phys. Rev. E*, 88(4):042809, 2013.

[87] L. HÉBERT-DUFRESNE : *La structure communautaire comme paradigme d'organisation des réseaux complexes*. Mémoire de maîtrise, Université Laval, 2011.

[88] 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(15):158702, 2011.

[89] 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(2):026108, 2012.

[90] L. HÉBERT-DUFRESNE, A. ALLARD, J.-G. YOUNG et L. J. DUBÉ : Universal growth constraints of human systems. *En préparation*

[91] L. HÉBERT-DUFRESNE, A. ALLARD, J.-G. YOUNG et L. J. DUBÉ : Global efficiency of local immunization on complex networks. *Sci. Rep.*, 3:2171, 2013.

[92] L. HÉBERT-DUFRESNE, A. ALLARD, J.-G. YOUNG et L. J. DUBÉ : Percolation on random networks with arbitrary k-core structure. *Phys. Rev. E*, 88:062820, 2013.

[93] L. Hébert-Dufresne, P.-A. Noël, V. Marceau, A. Allard et L. J. Dubé : Propagation dynamics on networks featuring complex topologies. *Phys. Rev. E*, 82(3):036115, 2010.

[94] L. Hébert-Dufresne, O. Patterson-Lomba, G. M. Goerg et B. M. Althouse : Pathogen mutation modeled by competition between site and bond percolation. *Phys. Rev. Lett.*, 110:108103, 2013.

[95] X. Huang, J. Gao, S. V. Buldyrev, S. Havlin et H. E. Stanley : Robustness of interdependent networks under targeted attack. *Phys. Rev. E*, 83(6):065101(R), 2011.

[96] T. Jia et A.-L. Barabási : Control capacity and a random sampling method in exploring controllability of complex networks. *Sci. Rep.*, 3:2354, 2013.

[97] T. Jia, Y.-Y. Liu, E. Csóka, M. Pósfai, J.-J. Slotine et A.-L. Barabási : Emergence of bimodality in controlling complex networks. *Nat. Commun.*, 4:2002, 2013.

[98] B. Karrer et M. E. J. Newman : Random graphs containing arbitrary distributions of subgraphs. *Phys. Rev. E*, 82(6):066118, 2010.

[99] M. J. Keeling et K. T. D. Eames : Networks and epidemic models. *J. R. Soc. Interface*, 2(4):295–307, 2005.

[100] M. J. Keeling et P. Rohani : *Modeling Infectious Diseases in Humans and Animals*. Princeton University Press, Princeton, NJ, 2007.

[101] A. K. Kel'mans : The reliability of information-transmission systems of arbitrary structure, taking into account the value of the information transmitted. *Avtomat. i Telemekh.*, 24:1250–9, 1963.

[102] E. Kenah et J. Robins : Second look at the spread of epidemics on networks. *Phys. Rev. E*, 76(3):036113, 2007.

[103] I. Z. Kiss et D. M. Green : Comment on "Properties of highly clustered networks". *Phys. Rev. E*, 78(4):048101, 2008.

[104] M. Kitsak, L. K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. E. Stanley et H. A. Makse : Identification of influential spreaders in complex networks. *Nat. Phys.*, 6:888–93, 2010.

[105] M. Kivelä, A. Arenas, M. Barthélemy, J. P. Gleeson, Y. Moreno et M. A. Porter : Multilayer Networks. *arXiv :1309.7233*.

[106] B. Klimt et Y. Yang : The Enron corpus : A new dataset for email classification research previous work on email classification. *Dans* J.-F. Boulicaut, F. Esposito, F. Giannotti et D. Pedreschi, éditeurs : *Mach. Learn. ECML 2004*, volume 3201 de *Lecture Notes in Computer Science*, pages 217–226. Springer Berlin Heidelberg, 2004.

[107] M. KURANT et P. THIRAN : Layered Complex Networks. *Phys. Rev. Lett.*, 96(13):138701, 2006.

[108] E. A. LEICHT et R. M. D'SOUZA : Percolation on interacting networks. *arXiv :0907.0894*, 2009.

[109] J. LESKOVEC, J. KLEINBERG et C. FALOUTSOS : Graph evolution : Densification and shrinking diameters. *ACM Trans. Knowl. Discov. Data*, 1(1):2, 2007.

[110] J. LESKOVEC, K. J. LANG, A. DASGUPTA et M. W. MAHONEY : Community structure in large networks : Natural cluster sizes and the absence of large well-defined clusters. *Internet Math.*, 6(1):29–123, 2009.

[111] Y.-Y. LIU, J.-J. SLOTINE et A.-L. BARABÁSI : Controllability of complex networks. *Nature*, 473(7346):167–73, 2011.

[112] Y.-Y. LIU, J.-J. SLOTINE et A.-L. BARABÁSI : Control centrality and hierarchical structure in complex networks. *PLoS One*, 7(9):e44459, 2012.

[113] Y.-Y. LIU, J.-J. SLOTINE et A.-L. BARABÁSI : Observability of complex systems. *Proc. Natl. Acad. Sci. USA*, 110(7):2460–5, 2013.

[114] K. MADDURI, D. EDIGER, K. JIANG, D. A. BADER et D. G. CHAVARRÍA-MIRANDA : A faster parallel algorithm and efficient multithreaded implementations for evaluating betweenness centrality on massive datasets. *Dans Proc. 3rd Work. Multithreaded Archit. Appl.*, 2009.

[115] S. MANIU, T. ABDESSALEM et B. CAUTIS : Casting a web of trust over Wikipedia : An interaction-based approach. *Dans Proc. 20th Int. Conf. Companion World wide web*, pages 87–88, New York, NY, USA, 2011. ACM.

[116] V. MARCEAU, P.-A. NOËL, L. HÉBERT-DUFRESNE, A. ALLARD et L. J. DUBÉ : Adaptive networks : Coevolution of disease and topology. *Phys. Rev. E*, 82(3):036116, 2010.

[117] V. MARCEAU, P.-A. NOËL, L. HÉBERT-DUFRESNE, A. ALLARD et L. J. DUBÉ : Modeling the dynamical interaction between epidemics on overlay networks. *Phys. Rev. E*, 84(2): 026105, 2011.

[118] M. MARDER : Dynamics of epidemics on random networks. *Phys. Rev. E*, 75(6):066103, 2007.

[119] S. MASLOV, K. SNEPPEN et A. ZALIZNYAK : Detection of topological patterns in complex networks : correlation profile of the internet. *Physica A*, 333:529–40, 2004.

[120] N. MASUDA : Immunization of networks with community structure. *New J. Phys.*, 11(12):123018, 2009.

[121] M. MEKATA : Kagome : The story of the basketweave lattice. *Phys. Today*, 56(2):12, 2003.

[122] 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(3):036112, 2011.

[123] C. D. MEYER : *Matrix Analysis and Applied Linear Algebra*. SIAM, 2001.

[124] L. A. MEYERS : Contact network epidemiology : Bond percolation applied to infectious disease prediction and control. *Bull. Amer. Math. Soc.*, 44(1):63–87, 2007.

[125] L. A. MEYERS, M. E. J. NEWMAN et B. POURBOHLOUL : Predicting epidemics on directed contact networks. *J. Theor. Biol.*, 240(3):400–18, 2006.

[126] L. A. MEYERS, B. POURBOHLOUL, M. E. J. NEWMAN, D. M. SKOWRONSKI et R. C. BRUNHAM : Network theory and SARS : Predicting outbreak diversity. *J. Theor. Biol.*, 232(1):71–81, 2005.

[127] J. C. MILLER : Percolation and epidemics in random clustered networks. *Phys. Rev. E*, 80(2):020901(R), 2009.

[128] R. MILO, S. SHEN-ORR, S. ITZKOVITZ, N. KASHTAN, D. CHKLOVSKII et U. ALON : Network motifs : simple building blocks of complex networks. *Science*, 298(5594):824–7, 2002.

[129] M. MOLLOY et B. REED : A critical point for random graphs with a given degree sequence. *Random Struct. Alg.*, 6:161–80, 1995.

[130] M. MOLLOY et B. REED : The size of the giant component of a random graph with a given degree sequence. *Comb. Probab. Comput.*, 7:295–305, 1998.

[131] J. MOODY : Race, school, and frienship segregation in America. *Am. J. Sociol.*, 107(3): 679–716, 2001.

[132] C. MOORE et M. E. J. NEWMAN : Exact solution of site and bond percolation on small-world networks. *Phys. Rev. E*, 62(5):7059–64, 2000.

[133] R. R. NADAKUDITI et M. E. J. NEWMAN : Graph spectra and the detectability of community structure in networks. *Phys. Rev. Lett.*, 108(18):188701, 2012.

[134] T. NEPUSZ et T. VICSEK : Controlling edge dynamics in complex networks. *Nat. Phys.*, 8(7):568–73, 2012.

[135] M. E. J. NEWMAN : Assortative mixing in networks. *Phys. Rev. Lett.*, 89(20):208701, 2002.

[136] M. E. J. NEWMAN : Spread of epidemic disease on networks. *Phys. Rev. E*, 66(1):016128, 2002.

[137] M. E. J. NEWMAN : Mixing patterns in networks. *Phys. Rev. E*, 67(2):026126, 2003.

[138] M. E. J. NEWMAN : Properties of highly clustered networks. *Phys. Rev. E*, 68(2):026121, 2003.

[139] M. E. J. NEWMAN : The structure and function of complex networks. *SIAM Rev.*, 45(2):167, 2003.

[140] M. E. J. NEWMAN : Power laws, Pareto distributions and Zipf's law. *Contemp. Phys.*, 46(5):323–51, 2005.

[141] M. E. J. NEWMAN : Modularity and community structure in networks. *Proc. Natl. Acad. Sci. USA*, 103(23):8577–82, 2006.

[142] M. E. J. NEWMAN : Component sizes in networks with arbitrary degree distributions. *Phys. Rev. E*, 76(4):045101(R), 2007.

[143] M. E. J. NEWMAN : Random graphs with clustering. *Phys. Rev. Lett.*, 103(5):058701, 2009.

[144] M. E. J. NEWMAN : *Networks : An Introduction*. Oxford University Press, 2010.

[145] M. E. J. NEWMAN et M. GIRVAN : Mixing patterns and community structure in networks. *Dans* R. PASTOR-SATORRAS, J. RUBI et A. DÍAZ-GUILERA, éditeurs : *Stat. Mech. Complex Networks*, pages 66–87. Springer, Berlin, 2003.

[146] M. E. J. NEWMAN et J. PARK : Why social networks are different from other types of networks. *Phys. Rev. E*, 68(3):036122, 2003.

[147] M. E. J. NEWMAN, S. H. STROGATZ et D. J. WATTS : Random graphs with arbitrary degree distributions and their applications. *Phys. Rev. E*, 64(2):026118, 2001.

[148] P.-A. NOËL : *Dynamiques stochastiques sur réseaux complexes*. Thèse de doctorat, Université Laval, 2012.

[149] P.-A. NOËL, A. ALLARD, L. HÉBERT-DUFRESNE, V. MARCEAU et L. J. DUBÉ : Propagation on networks : An exact alternative perspective. *Phys. Rev. E*, 85(3):031118, 2012.

[150] P.-A. NOËL, A. ALLARD, L. HÉBERT-DUFRESNE, V. MARCEAU et L. J. DUBÉ : Spreading dynamics on complex networks : a general stochastic approach. *J. Math. Biol.*, 2013.

[151] P.-A. NOËL, B. DAVOUDI, R. C. BRUNHAM, L. J. DUBÉ et B. POURBOHLOUL : Time evolution of epidemic disease on finite and infinite networks. *Phys. Rev. E*, 79(2):026101, 2009.

[152] K. PAECH, W. BAUER et S. PRATT : Zipf's law in nuclear multifragmentation and percolation theory. *Phys. Rev. C*, 76(5):054603, 2007.

[153] G. PALLA, I. DERÉNYI, I. FARKAS et T. VICSEK : Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–8, 2005.

[154] G. Palla, I. J. Farkas, P. Pollner, I. Derényi et T. Vicsek : Directed network modules. *New J. Phys.*, 9(6):186, 2007.

[155] G. Palla, I. J. Farkas, P. Pollner, I. Derényi et T. Vicsek : Fundamental statistical features and self-similar properties of tagged networks. *New J. Phys.*, 10(12):123026, 2008.

[156] J. Park et M. E. J. Newman : Origin of degree correlations in the Internet and other networks. *Phys. Rev. E*, 68(2):026112, 2003.

[157] R. Pastor-Satorras et A. Vespignani : Epidemic spreading in scale-free networks. *Phys. Rev. Lett.*, 86(14):3200–3, 2001.

[158] R. Pastor-Satorras et A. Vespignani : Immunization of complex networks. *Phys. Rev. E*, 65(3):036104, 2002.

[159] B. Pittel : On the probable behaviour of some algorithms for finding the stability number of a graph. *Math. Proc. Cambridge Philos. Soc.*, 92:511–26, 1982.

[160] M. Pósfai, Y.-Y. Liu, J.-J. Slotine et A.-L. Barabási : Effect of correlations on network controllability. *Sci. Rep.*, 3:01067, 2013.

[161] W. H. Press, S. A. Teukolsky, W. T. Vetterling et B. P. Flannery : *Numerical Recipes : The Art of Scientific Computing*. Cambridge University Press, New York, NY, USA, 3^{e} édition, 2007.

[162] E. Ravasz et A.-L. Barabási : Hierarchical organization in complex networks. *Phys. Rev. E*, 67(2):026112, 2003.

[163] F. Reif : *Fundamentals of Statistical and Thermal Physics*. McGraw-Hill, 1965.

[164] P. J. Reynolds, W. Klein et H. E. Stanley : A real-space renormalization group for site and bond percolation. *J. Phys. C*, 10(8):L167–L172, 1977.

[165] P. J. Reynolds, H. E. Stanley et W. Klein : Large-cell Monte Carlo renormalization group for percolation. *Phys. Rev. B*, 21(3):1223–1245, 1980.

[166] M. Ripeanu et I. Foster : Mapping the Gnutella network : Macroscopic properties of large-scale peer-to-peer systems. *In* P. Druschel, F. Kaashoek et A. Rowstron, éditeurs : *Peer-to-Peer Syst.*, pages 85–93. Springer, Berlin & Heidelberg, 2002.

[167] A. Saadatpour, R.-S. Wang, A. Liao, X. Liu, T. P. Loughran, I. Albert et R. Albert : Dynamical and structural analysis of a T cell survival network identifies novel candidate therapeutic targets for large granular lymphocyte leukemia. *PLoS Comput. Biol.*, 7(11):e1002267, 2011.

[168] S. Sahasrabudhe et A. E. Motter : Rescuing ecosystems from extinction cascades through compensatory perturbations. *Nat. Commun.*, 2(1):170, 2011.

[169] M. SALATHÉ et J. H. JONES : Dynamics and control of diseases in networks with community structure. *PLoS Comput. Biol.*, 6(4):e1000736, 2010.

[170] D. J. SALKELD, M. SALATHÉ, P. STAPP et J. H. JONES : Plague outbreaks in prairie dog populations explained by percolation thresholds of alternate host abundance. *Proc. Natl. Acad. Sci. USA*, 107(32):14247–50, 2010.

[171] C. SCHMELTZER, J. SORIANO, I. M. SOKOLOV et S. RÜDIGER : Percolation of spatially constrained Erdős-Rényi networks with degree correlations. *Phys. Rev. E*, 89(1):012116, 2014.

[172] C. R. SCULLARD et R. M. ZIFF : Critical surfaces for general bond percolation problems. *Phys. Rev. Lett.*, 100(18):185701, 2008.

[173] C. R. SCULLARD et R. M. ZIFF : Critical surfaces for general inhomogeneous bond percolation problems. *J. Stat. Mech.*, 2010(3):P03021, 2010.

[174] S. B. SEIDMAN : Network structure and minimum degree. *Soc. Networks*, 5:269–87, 1983.

[175] M. Á. SERRANO et M. BOGUÑÁ : Clustering in complex networks. I. General formalism. *Phys. Rev. E*, 74(5):056114, 2006.

[176] M. Á. SERRANO et M. BOGUÑÁ : Clustering in complex networks. II. Percolation properties. *Phys. Rev. E*, 74(5):056115, 2006.

[177] M. Á. SERRANO et M. BOGUÑÁ : Percolation and epidemic thresholds in clustered networks. *Phys. Rev. Lett.*, 97(8):088701, 2006.

[178] M. Á. SERRANO, D. KRIOUKOV et M. BOGUÑÁ : Self-similarity of complex networks and hidden metric spaces. *Phys. Rev. Lett.*, 100(7):078701, 2008.

[179] V. K. S. SHANTE et S. KIRKPATRICK : An introduction to percolation theory. *Adv. Phys.*, 20(85):325–57, 1971.

[180] X. SHI, L. A. ADAMIC et M. J. STRAUSS : Networks of strong ties. *Physica A*, 378(1):33–47, 2007.

[181] S.-W. SON, G. BIZHANI, C. CHRISTENSEN, P. GRASSBERGER et M. PACZUSKI : Percolation theory on interdependent networks based on epidemic spreading. *EPL*, 97(1):16006, 2012.

[182] D. SORNETTE : *Critical Phenomena in Natural Sciences*. Springer, 2004.

[183] C. SPEARMAN : The proof and measurement of association between two things. *Amer. J. Psychol.*, 15(1):72–101, 1904.

[184] M. R. SPIEGEL : *Theory and Problems of Probability and Statistics*. McGraw-Hill, New York, NY, USA, 1992.

[185] H. E. STANLEY : *Introduction to Phase Transitions and Critical Phenomena*. Oxford University Press, Oxford, UK, 1971.

[186] D. STAUFFER et A. AHARONY : *Introduction to percolation theory*. Taylor and Francis, London, 2^{e} édition, 1994.

[187] Walter H. STOCKMAYER : Theory of Molecular Size Distribution and Gel Formation in Branched-Chain Polymers. *J. Chem. Phys.*, 11(2):45, 1943.

[188] S. H. STROGATZ : *Nonlinear Dynamics and Chaos : With Applications to Physics, Biology, Chemistry, and Engineering*. Perseus Books Group, 1994.

[189] J. SUN et A. E. MOTTER : Controllability transition and nonlocality in network control. *Phys. Rev. Lett.*, 110(20):208701, 2013.

[190] M. F. SYKES et J. W. ESSAM : Exact critical percolation probabilities for site and bond problems in two dimensions. *J. Math. Phys.*, 5(8):1117, 1964.

[191] W. TRAUTMANN : Multifragmentation and the liquid-gas phase transition : An experimental overview. *Nucl. Phys. A*, 752:407c–416c, 2005.

[192] A. VAZQUEZ : Spreading dynamics on heterogeneous populations : Multitype network approach. *Phys. Rev. E*, 74(6):066114, 2006.

[193] A. VÁZQUEZ et Y. MORENO : Resilience to damage of graphs with degree correlations. *Phys. Rev. E*, 67(1):015101(R), 2003.

[194] A. VÁZQUEZ, R. PASTOR-SATORRAS et A. VESPIGNANI : Large-scale topological and dynamical properties of the Internet. *Phys. Rev. E*, 65(6):066130, 2002.

[195] A. VESPIGNANI : The fragility of interdependency. *Nature*, 464:984–985, 2010.

[196] B. VISWANATH, A. MISLOVE, M. CHA et K. P. GUMMADI : On the evolution of user interaction in Facebook. *Dans Proc. 2nd ACM Work. Online Soc. networks - WOSN '09*, pages 37–42, 2009.

[197] D. J. WATTS et S. H. STROGATZ : Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–2, 1998.

[198] D. B. WEST : *Introduction to Graph Theory*. Prentice-Hall, Upper Saddle River, NJ, 2^{e} édition, 2001.

[199] G. YAN, J. REN, Y.-C. LAI, C.-H. LAI et B. LI : Controlling complex networks : How much energy is needed ? *Phys. Rev. Lett.*, 108(21):218703, 2012.

[200] Y. YANG, J. WANG et A. E. MOTTER : Network observability transitions. *Phys. Rev. Lett.*, 109(25):258701, 2012.

[201] R. D. ZIMMERMAN, C. E. MURILLO-SÁNCHEZ et R. J. THOMAS : MATPOWER : Steady-state operations, systems research and education. *IEEE Trans. Power Syst.*, 26(1):12–9, 2011.

[202] V. ZLATIĆ, D. GARLASCHELLI et G. CALDARELLI : Networks with arbitrary edge multiplicities. *EPL*, 97(2):28005, 2012.