LAURENT HÉBERT-DUFRESNE

# LA STRUCTURE COMMUNAUTAIRE COMME PARADIGME D'ORGANISATION DES RÉSEAUX COMPLEXES

Mémoire présenté
à la Faculté des études supérieures de l'Université Laval
dans le cadre du programme de maîtrise en physique
pour l'obtention du grade de Maître ès sciences (M.Sc.)

DÉPARTEMENT DE PHYSIQUE, DE GÉNIE PHYSIQUE ET D'OPTIQUE
FACULTÉ DES SCIENCES ET DE GÉNIE
UNIVERSITÉ LAVAL
QUÉBEC

2011

© Laurent Hébert-Dufresne, 2011

# Résumé

La caractérisation des *propriétés universelles* des systèmes complexes aide à comprendre *comment* leurs éléments sont distribués et connectés, *pourquoi* il en est ainsi et quelles en sont les conséquences. Dans les réseaux complexes, ces propriétés incluent l'organisation *indépendente d'échelle*, la propriété du *small-world*, la *modularité* et l'*auto-similarité*. Par contre, aucun mécanisme connu n'explique l'émergence de toutes ces propriétés. On développe ici un nouveau modèle d'organisation qui considère les communautés, plutôt que les éléments de base ou les liens qu'ils partagent, comme les blocs fondamentaux des systèmes complexes. On conclut que les propriétés mentionnées précédemment sont unifiées dans une structure communautaire indépendente d'échelle. Comme preuve empirique de notre *attachement préférentiel structurel*, nous examinons des réseaux sociaux (collaborations entre scientifiques et entre acteurs) et d'information (l'Internet) et sommes capables de reproduire leurs distributions en éléments par communauté et communautés par élément. De plus, notre modèle permet de prédire comment les structures et les éléments sont interconnectés, souvent de manière auto-similaire, en plus d'offrir de bons indices quant à l'évolution temporelle de ces systèmes.

Nous utilisons la structure communautaire indépendente d'échelle comme nouveau paradigme d'organisation et étudions ces effets sur les *phénomènes propagatoires* sur réseaux complexes. Ici, une analyse par champs moyens est utilisée pour coupler de façon cohérente la dynamique des éléments du réseau et la dynamique des motifs récurrents dans leur topologie. Pour un modèle d'épidémie sur réseaux sociaux, cette approche procure un système d'EDOs pour l'évolution temporelle, en plus de solutions analytiques pour le seuil épidémique et la prévalence à l'équilibre. Dans le cadre de cette application, nous évaluons comment notre compréhension de la structure d'un réseau nous aide à contrôler sa dynamique. À la lumière de nos analyses, nous postulons ensuite un nouveau paradigme pour la description de l'*organisation des réseaux complexes*.

# Abstract

Characterizing *universal properties* of complex systems provides insights on *how* elements are distributed and connected in nature, which in turn might help understand *why* things follow any specific design and *what* are its consequences. In complex networks, these properties include *scale-free design*, the *small-world* property, *modularity* and *self-similarity*. However, no known mechanism explains the emergence of all these properties. We develop a new organization model by considering that communities are the fundamental building blocks of complex systems, instead of the basic elements or the links they share. We find that the aforementioned properties are encompassed within a scale-free community structure. As testing ground for our *structural preferential attachment*, we examine social (scientific and actor collaborations) and information (Internet) networks, and are able to reproduce their distributions of elements per community and of communities per element. More interestingly, our approach also predicts how the structures and elements are interconnected, often in a self-similar manner, and even provides insights into the time evolution of such systems.

We use this scale-free community structure as a new paradigm for network organization and investigate its effects on *propagation phenomena* on complex networks. Here, a mean-field analysis is used to coherently couple the dynamics of the network elements on the one hand and their recurrent topological patterns on the other hand. In a model of epidemic spread on social networks, this approach yields a set of ODEs for the time evolution of the system, as well as analytical solutions for the epidemic threshold and equilibria. In the context of this particular application, we discuss how our understanding of a network structure helps us control its dynamics. In light of our analysis, we finally postulate a new paradigm for the description of *complex network organization*.

# Avant-propos

Comme vous serez à même de le constater, ce mémoire tient autant du carnet de voyage que de la thèse scientifique.

Le premier chapitre constitue une sorte de résumé des différents concepts et lectures qui ont su frapper mon imaginaire lorsque j'ai entrepris mon exploration des réseaux complexes. Bien que tous les résultats présentés aient été produits pour ce mémoire, ils sont en fait reproduits ou inspirés de la littérature. Les trois chapitres suivants sont des retranscriptions de trois des articles que j'ai eu le plaisir de rédiger pendant ma maîtrise. À noter que les articles vous seront présentés ici dans le désordre et accompagnés de mises en contexte ; dans l'espoir de raconter une histoire plus passionnante. En tant que premier auteur, j'ai été initiateur et responsable de chaque étape des travaux qui y sont détaillés. Mis à part quelques changements dans la mise en page et une modification mineure au texte, ils sont présentés tel que publiés dans leur journal respectif. À l'exception du troisième article, qui se trouve au Chapitre 4 et qui est en fait le premier que j'ai écrit, où une erreur qui avait réussi à se glisser dans la version finale est ici corrigée. Le chapitre de conclusion jette pour sa part un dernier regard sur le chemin parcouru avant de se tourner vers le futur.

Cela étant dit, je crois que c'est le bon endroit pour faire mention de ceux sans qui je n'aurais pu trouver l'inspiration, les connaissances ou le plaisir pour mener mes projets à terme. Il est en fait approprié que ce projet, axé sur les systèmes où l'ensemble est plus complexe que la somme des éléments, n'aurait lui-même jamais pu voir le jour sans le bon regroupement de gens passionnés.

En premier lieu, je me dois d'abord de remercier mon directeur Louis J. Dubé, principalement pour son ouverture d'esprit essentielle à un projet en constante métamorphose. Au-delà de son partage de connaissances théoriques, d'histoires et de contrepèteries qui ne se racontent pas ici (ou nulle part d'ailleurs), il a surtout favorisé ce projet en poussant ses étudiants au meilleur d'eux-mêmes peu importe la route qu'ils choisissaient d'emprunter.

En second lieu, une mention particulière est due à mes collègues de tous les jours : Antoine Allard, Vincent Marceau et Pierre-André Noël. C'est particulièrement plaisant de délirer

autour d'un café ou d'une bière lorsque, de temps à autre, une idée géniale semble sortir de nulle part. Je ne peux imaginer de meilleurs coéquipiers pour faire de la recherche scientifique de manière rigoureuse, créative et amusante.

En troisième lieu, je désire remercier le reste du groupe de recherche Dynamica, Denis Gagnon, Guillaume Painchaud-April, Julien Poirier et Jean-Gabriel Young, pour les conseils de certains, le cynisme d'un autre, et la passion de tous. Les pauses cafés et les soirées interminables ne semblent jamais assez longues pour épuisez vos idées. Que cela continue.

Une mention bien particulière est due aux professeurs Alain Hertz et Pierre Mathieu pour avoir accepté d'examiner ce mémoire de maîtrise. Mon projet de recherche étant essentiellement à cheval entre leur expertise respective, je n'aurais pu espérer voir mon travail examiné par de meilleurs regards extérieurs que les leurs.

La fin de cet Avant-propos approche, mais je dois un remerciement bien spécial à la première personne à me venir en tête lorsque je repense à ces deux dernières années : ma rose des sables, Meriem. Ton amour me permet d'être heureux quand ça va moins bien ; ton humour m'empêche de me prendre trop au sérieux ; alors que ta complexité et ton intelligence m'étonnent à chaque jour. Tu mérites bien ton nom dans ce mémoire.

J'aimerais aussi mentionner les organismes qui ont apporté leur soutien financier à mes études, d'abord mes parents, Guy Dufresne et Danielle Hébert, puis le Conseil de recherche en sciences naturelles et en génie (CRSNG), les Instituts de recherche en santé du Canada (IRSC) et le Fond québécois de recherche sur la nature et les technologies (FQRNT).

Bref, merci à mes parents, à ma soeur Lysiane, à mes amis et à Meriem pour leur énergie et leur soutien.

Et merci à mes collègues pour leurs idées et leur passion. Ceux qui s'en vont, ce fut un plaisir. Ceux qui restent, je ne sais pas ce qu'on va faire, mais on va le faire en *******!

*À Meriem, à la folie...*

*"Men, it has been well said, think in herds ; it will be seen that they go mad in herds, while they only recover their senses slowly, and one by one."*
*— Charles Mackay (1814 – 1889) dans Extraordinary Popular Delusions and the Madness of Crowds*

# Table des matières

**Résumé** ii

**Abstract** iv

**Avant-Propos** vi

**Table des matières** xiii

**Liste des figures** xvi

**Liste des tableaux** xvii

**Liste des abbréviations** xvii

**Liste des contributions** xix

**Prologue** 1

**1 Réseaux complexes : questions d'universalité et de connexions** 5

1.1 Introduction aux réseaux . . . . . 5

1.1.1 Königsberg, 1735 . . . . . 5

1.1.2 La structure d'un réseau . . . . . 6

1.1.3 La physique statistique et les réseaux complexes . . . . . 8

1.1.4 Les réseaux Erdős–Rényi . . . . . 9

1.2 Réseaux libres d'échelle . . . . . 11

1.2.1 Barabási et Albert, 1999 . . . . . 12

1.2.2 Conséquences de l'indépendance d'échelle . . . . . 13

1.2.3 Modèle canonique de réseaux aléatoires . . . . . 16

1.3 Réseaux modulaires et les *small-worlds* . . . . . 17

1.3.1 Watts et Strogatz, 1998 . . . . . 17

1.3.2 La structure communautaire . . . . . 20

1.4 L'auto-similarité des réseaux complexes . . . . . 24

1.4.1 Song *et al.*, 2005 . . . . . 24

1.5 Propriétés universelles des réseaux complexes . . . . . 25

1.5.1 Vérification et application : l'épidémiologie sur réseaux . . . . . 26
1.5.2 Vers une description simplifiée . . . . . 27

**2 L'attachement préférentiel structurel : les réseaux au-delà du lien 29**
2.1 Avant-propos . . . . . 30
2.2 Résumé . . . . . 30
2.3 Abstract . . . . . 30
2.4 A universal matter. . . . . 30
2.5 On the matter of networks. . . . . 31
2.6 When structure matters. . . . . 32
2.7 A simple model. . . . . 33
2.8 Results and discussions. . . . . 35
2.9 Conclusion and perspective. . . . . 37
2.10 Appendix : Excerpts of Supplemental Material . . . . . 38
2.10.1 Data . . . . . 38
2.10.2 Levels of organization . . . . . 39
2.10.3 A note on node-based and link-based systems . . . . . 39
2.10.4 Supplementary results and discussions . . . . . 39
2.10.5 From communities, back to links . . . . . 41

**3 Croissance de systèmes libres d'échelle, modulaires et auto-similaires 43**
3.1 Avant-propos . . . . . 44
3.2 Résumé . . . . . 44
3.3 Abstract . . . . . 44
3.4 Introduction . . . . . 45
3.5 Stochastic process . . . . . 47
3.5.1 Time evolution . . . . . 47
3.5.2 Degree distributions . . . . . 48
3.5.3 Statistical equilibrium . . . . . 50
3.5.4 Scaling behaviour . . . . . 51
3.6 Approximations and limitations . . . . . 52
3.6.1 Correspondence between system bases . . . . . 52
3.6.2 Multiple memberships, multiple links and self-loops . . . . . 54
3.6.3 Element-structure correlations . . . . . 55
3.7 Peloton dynamics . . . . . 57
3.8 Conclusion . . . . . 59
3.9 Appendix : Explicit solution to continuous time SPA . . . . . 61
3.9.1 Definition of a continuous time PA process . . . . . 61
3.9.2 Explicit solution . . . . . 62

**4 Dynamique de propagation sur structure communautaire 65**

4.1 Avant-propos . . . . . 66
4.2 Résumé . . . . . 66
4.3 Abstract . . . . . 66
4.4 Introduction . . . . . 67
4.5 Community structure . . . . . 68
4.6 SIS model of epidemics on community structure . . . . . 70
4.6.1 Construction of the dynamical model . . . . . 70
4.6.2 Solution for network stable state . . . . . 73
4.6.3 Solution for epidemic threshold . . . . . 74
4.7 Implementation and validation . . . . . 75
4.7.1 Treatment of the analytical model . . . . . 75
4.7.2 Numerical model . . . . . 77
4.7.3 Results on Newman's topology . . . . . 78
4.7.4 Results on a general topology . . . . . 80
4.8 Conclusion . . . . . 80
4.9 Appendix A : Community structure, without degree correlation, raises the epidemic threshold . . . . . 81

**5 Conclusion et perspectives : un nouveau paradigme** **85**

**Bibliographie** **89**

# Liste des figures


1.1 Les ponts de Königsberg. . . . . 6
1.2 Réseaux dirigés et pondérés . . . . . 7
1.3 Les Misérables. . . . . 9
1.4 L'urbaniste aléatoire. . . . . 10
1.5 Distribution en degré du réseau de Königsberg . . . . . 11
1.6 Réseaux libres d'échelle . . . . . 12
1.7 Pannes et attaques sur l'Internet . . . . . 15
1.8 Réseaux ordonnés et désordonnés . . . . . 19
1.9 Réseaux *small-world* : entre l'ordre et le désordre . . . . . 20
1.10 Motifs de réseaux plus ou moins ordonnés . . . . . 21
1.11 Exemple de réseau à structure communautaire. . . . . 21
1.12 Détection des complexes de protéines de la levure *S. Cerevisiae*. . . . . 23
1.13 Auto-similarité des niveaux d'organisation du *cond-mat arXiv* . . . . . 25
1.14 Épidémie de grippe sur *arXiv* . . . . . 27

2.1 Spectrum of scale-free complex systems . . . . . 32
2.2 Structural preferential attachment and the systems it creates . . . . . 34
2.3 Reproduction of real systems with SPA . . . . . 36
2.4 Supplementary results . . . . . 40

3.1 A step of node-based SPA . . . . . 46
3.2 Validation of the analytical description of SPA . . . . . 49
3.3 Validation of the asymptotic solution of SPA . . . . . 51
3.4 Validation of the scaling exponents of SPA . . . . . 52
3.5 Comparison between node-based and link-based SPA . . . . . 53
3.6 Reproducing the arXiv with analytical results . . . . . 53
3.7 Finite size effects in SPA . . . . . 55
3.8 Element-structure correlations in SPA and in the arXiv . . . . . 56
3.9 Theoretical observation of peloton dynamics . . . . . 58
3.10 Peloton dynamics in prose samples . . . . . 60

4.1 Community structure : topology and formalism . . . . . 68
4.2 Identifying the epidemic threshold and equilibrium of a disease . . . . . 74

4.3 Degree distributions of two community structured system . . . . . 76
4.4 Time evolution and steady state of an epidemic on community structure . . . 76
4.5 Time evolution and steady state of an epidemic on community structure 2 . . 78

5.1 Simulation de quarantaine sur un réseau social . . . . . 86
5.2 Prédiction d'une épidémie sur *arXiv* . . . . . 87

# Liste des tableaux

1.1 Résultats de Watts et Strogatz . . . . . 20

2.1 Scaling exponents of SPA . . . . . 35

3.1 Parameters of a general preferential attachment process . . . . . 62

# Liste des abbréviations

## Abbréviations françaises

| | |
|---|---|
| APS | Attachement Préférentiel Structurel |
| EDO | Équation Différentielle Ordinaire |
| ER | Réseau Erdős–Rényi |
| RAE | Réseau Aléatoire Équivalent |
| SA | Système Autonome |
| SIS | Susceptible-Infectieux-Susceptible |

## Abbréviations anglaises

| | |
|---|---|
| BA | Barabási-Albert model |
| CS | Community Structure |
| ERN | Equivalent Random Network |
| GF | Generating Function |
| IMDb | Internet Movie Database |
| MC | Monte Carlo |
| ODE | Ordinary Differential Equation |
| PA | Preferential Attachment |
| PGF | Probability Generating Function |
| RN | Random Network |
| SIS | Susceptible-Infectious-Susceptible |
| SPA | Structural Preferential Attachment |

# Liste des contributions

- 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 (2010), 036115.
- L. Hébert-Dufresne, A. Allard, V. Marceau, P.-A. Noël, et L.J. Dubé, *Structural preferential attachment : Network organization beyond the link*, Phys. Rev. Lett. 107 (2011), 158702.
- 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*, arXiv :1109.0034 (en soumission).
- L. Hébert-Dufresne, V. Marceau, P.-A. Noël, A. Allard, et L.J. Dubé, *The Social Zombie : modelling undead outbreaks on social networks*, *dans* Mathematical Modelling of Zombies. Robert Smith ?, ed. University of Ottawa Press, (à paraître en 2012).
- 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 (2010), 036116.
- 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 (2011), 026105.
- P.-A. Noël, A. Allard, L. Hébert-Dufresne, V. Marceau, et L.J. Dubé, *Propagation on networks : an exact alternative perspective*, arXiv :1102.0987 (en soumission).

# Prologue

## Ouverture

L'être humain, vu comme système dynamique, est assez simple. L'apparente complexité de notre comportement au fil du temps est principalement une réflexion de la complexité de l'environnement dans lequel nous nous trouvons. Cette pensée, que nous devons à Herbert A. Simon dans *Models of Man* [73], fut et sera la principale motivation pour le projet de recherche qui sera ici exposé.

Car l'Homme, à bien des égards, est si complexe que tous les systèmes généralement étudiés par la physique pâlissent sous comparaison. Après tout, nos actions sont guidées non pas par des lois mathématiques, mais par des forces abstraites de morale, d'amour, de haine et de désirs. Ces forces sont si complexes que la psychologie d'un individu peut mystifier à elle seule le plus brillant des psychanalystes. Alors, dans quel contexte peut-on penser que l'être humain est un objet simple ?

Considérons un instant le fait qu'il n'est pas nécessaire, ou même souhaitable, de partir d'une description en termes des quarks et de leurs interactions pour formuler les lois de la thermodynamique, car la physique statistique nous offre des outils pour décrire les comportements globaux ou *macroscopiques* de larges quantités d'éléments sans considérer toute la complexité de leur nature ou de leurs interactions *microscopiques*. En partant de cette philosophie, on peut avoir bon espoir de décrire certaines propriétés et comportements de larges systèmes sociaux sans se perdre dans la complexité de chaque sujet. On verra également que cette approche n'est en rien limitée aux humains, et peut être utilisée pour décrire des systèmes de toute nature où de nombreux éléments se trouvent à former une large toile d'interactions.

Ainsi, bien que les systèmes étudiés dans ce mémoire soient incroyablement variés, de l'Internet à Hollywood, ils partagent tous une caractéristique fondamentale : ils sont réductibles à un ensemble de points et de lignes. Les points peuvent bien représenter des systèmes autonomes de l'Internet ou des acteurs hollywoodiens et les lignes un ensemble de routage ou

des collaborations professionnelles, la base du système reste un simple *réseau* de connexions. Que ce soit en se rendant à son travail (réseaux routiers et de transport en commun) ou en essayant d'en trouver un (réseaux de contacts), en téléphonant (réseaux de satellites) ou même simplement en allumant la radio (réseaux électriques et d'information), ces systèmes nous entourent en permanence. En améliorant notre compréhension de leur structure, on améliore évidemment notre habileté à les utiliser à notre avantage. Et pour bien les comprendre, on doit les reproduire numériquement, les décrire mathématiquement, mais avant tout, les modéliser théoriquement.

Le travail du modélisateur est alors de balancer la quantité d'information du système d'origine avec la généralité de son modèle. Évidemment, on ne veut pas devoir réinventer la roue à chaque fois que l'on se trouve confronté à un nouveau réseau. Il s'agit de simplifier le comportement des éléments du système pour que la complexité du problème réside dans l'organisation de leurs interactions et non dans leur nature. Le réseau résultant est alors suffisant pour modéliser de multiples propriétés et processus dynamiques reliés au système en question : robustesse face aux défaillances, vulnérabilité aux attaques ou comportement de phénomènes propagatoires (de virus, savoir ou autres).

## La science des réseaux

La science des réseaux est l'étude d'ensembles d'éléments et de liens représentant des phénomènes physiques, biologiques ou sociaux menant à des modèles prédictifs. Telle que définie ici, cette science est relativement nouvelle et éloignée des sujets typiques de recherche en physique théorique. Cependant, cette conception est en voie de disparaître en raison d'un mouvement de masse de physiciens issus de la physique statistique vers la modélisation de ces systèmes. En fait, l'effervescence qu'a connu la science des réseaux au cours des vingt dernières années est comparable à l'explosion de l'étude du chaos au XXe siècle. Cette comparaison est particulièrement intéressante considérant que, comme la théorie du chaos, la science des réseaux est l'un des paradigmes de la complexité. La première s'intéresse typiquement aux systèmes simples à comportements complexes (e.g. un pendule forcé qui s'emballe), alors que la seconde étudie des systèmes complexes soumis à des lois simples (e.g. une information qui se propage sur une toile de liens).

Un des fers de lance de la science des réseaux est l'étude des processus de croissance ou d'organisation. Ces processus tentent de modéliser l'émergence des propriétés fondamentales des réseaux réels en liant les éléments d'un réseau selon des lois simples et indépendantes de la nature du système. En effet, plusieurs propriétés se retrouvent autant dans les réseaux sociaux ou neuronaux que dans les toiles alimentaires, les interactions de protéines, ou le World Wide Web. L'attrait de ces recherches est double. D'une part, elles permettent de rassembler

les systèmes naturels en classes distinctes, de sorte qu'une conclusion tirée sur un système donné peut être utilisée pour l'étude d'autres systèmes partageant les mêmes propriétés fondamentales et appartenant donc à la même classe. D'autre part, elles permettent une meilleure compréhension des règles qui dictent la croissance de réseaux naturels et simplifient ainsi considérablement les comportements pourtant complexes des éléments du système.

## Un nouveau paradigme

Par le présent mémoire, on cherche à poser une question bien précise. Pourquoi accorder autant d'importance aux liens ? Il existe d'autres façons d'organiser un ensemble d'éléments qu'en les reliant un à un et par conséquent, d'autres façons de concevoir des réseaux. Par exemple, demandez à un individu combien d'amis il possède, et vous aurez droit à une réponse floue, au mieux un ordre de grandeur. Demandez-lui plutôt à combien de groupes sociaux il appartient, par exemple sa famille, son travail, ses loisirs, et vous aurez alors une réponse beaucoup plus précise.

On propose ainsi d'étendre le concept de liens pour constater que, bien qu'ils constituent l'unité de base des interactions dans le système, ils ne sont pas nécessairement le niveau d'organisation le plus significatif. En changeant notre façon de voir les réseaux complexes, on en apprendra davantage sur leurs propriétés et on observera ce que l'on connaît déjà sous un tout autre jour.

L'ouvrage est divisé comme suit. Au chapitre 1, on définit les fondements de l'étude des réseaux complexes, puis on présente leurs propriétés *universelles*. Au chapitre 2, on introduit un nouveau processus stochastique, *l'attachement préférentiel structurel*. Ce processus reproduit les propriétés statistiques des topologies de divers réseaux réels et unifie les propriétés universelles présentées au chapitre 1. Cet attachement préférentiel structurel est étudié en détail au cours du chapitre 3 ; divers outils de la physique statistique sont mis en oeuvre pour illustrer les points forts et les points faibles du modèle proposé. Le chapitre 4 est consacré au développement d'un formalisme dynamique décrivant l'évolution temporelle d'un processus de propagation d'épidémies sur des réseaux possédant une topologie non-aléatoire (notons au passage que l'ordre n'est généralement pas synonyme de simplicité en science des réseaux). En conclusion, nous extrapolons à partir du travail accompli pour paver la voie vers un nouveau paradigme d'organisation des réseaux complexes.

# Chapitre 1

# Réseaux complexes : questions d'universalité et de connexions

Au sens large, un réseau est un ensemble d'éléments potentiellement connectés. On nomme les éléments des *noeuds* et les connexions des *liens*.

## 1.1 Introduction aux réseaux

L'étude des objets mathématiques qui représentent des réseaux, les *graphes*, est beaucoup plus ancienne que la science des réseaux telle que décrite en ouverture. Il est pratique, et pédagogique, de revenir à la naissance de cette branche des mathématiques pour ce faire une idée concrète de ce que sont les réseaux.

### 1.1.1 Königsberg, 1735

C'est en 1735 que Leonhard Euler utilisa le concept de graphe pour résoudre le problème des sept ponts de Königsberg [18]. La ville avait en effet sept ponts traversant la rivière Pregel et reliant la terre ferme à deux îles. La question était de savoir s'il était possible de traverser chaque pont qu'une seule fois sans jamais revenir sur ses pas. La solution ingénieuse d'Euler était la suivante : la configuration des îles et du continent n'importe pas dans le problème, on peut donc les représenter par des simples points (des *noeuds*), toute l'information pertinente réside plutôt dans les sept ponts (des *liens*) entre ces points. Euler se retrouve alors avec un graphe (un *réseau*) dont l'étude des propriétés topologiques permet de solutionner le problème (voir Fig. 1.1). Dans le cas d'espèce, un parcours comme celui demandé n'existe que s'il y a au maximum deux noeuds ayant un *degré* impair (i.e. possédant un nombre impair

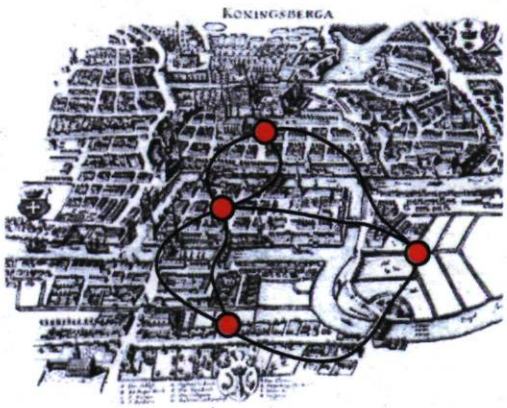



FIG. *1.1:* **Les ponts de Königsberg.** Carte de Königsberg du vivant d'Euler avec, en superposition, une schématisation de la structure du réseau sous-jacent. Chaque morceau de terre est un noeud et chaque pont un lien.

de liens) : la réponse à la question était donc négative.[1]

La solution d'Euler, aussi simple soit-elle, représente les premiers balbutiements d'un nouvel outil d'analyse, d'un nouveau champ d'étude, voire d'une nouvelle science. En fait, ce problème est pratiquement suffisant pour illustrer presque toutes les propriétés de base d'un réseau : la quantité de noeuds (quatre pour le problème des ponts de Königsberg), la quantité de liens (sept) et le concept de degré d'un noeud (son nombre de liens). Ce n'est pas assez d'information pour se faire une idée précise du réseau en question, mais c'est un début.

### 1.1.2 La structure d'un réseau

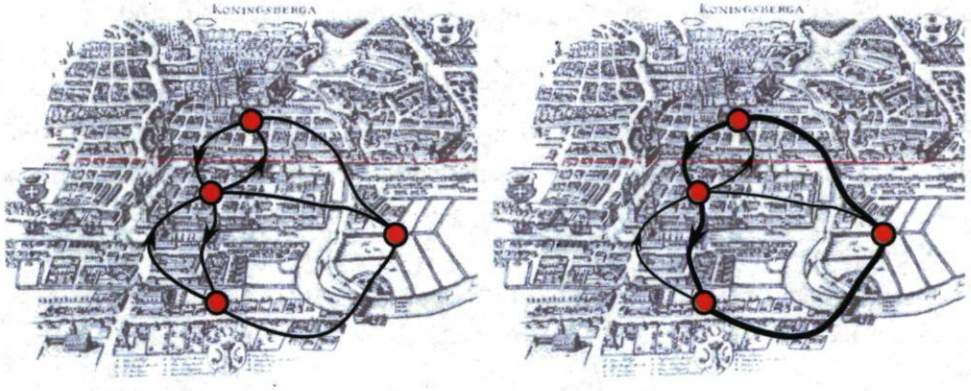
Comment représenter de façon exacte et unique le réseau illustré sur la Fig. 1.1 ? Comment pouvons nous écrire toute l'information nécessaire à sa construction ? Nous pourrions compter le nombre de triangles (ensemble de trois noeuds connectés : deux), de liens multiples (deux liens doubles) et d'autres structures particulières. Mais comment s'assurer qu'il ne reste aucun arbitraire dans le choix de *qui est connecté avec qui* ? La réponse est donc de littéralement écrire le réseau sous forme mathématique en utilisant une matrice. Pour un réseau possédant $N$ noeuds, on aura une matrice $\mathbf{A}$ de taille $N \times N$ telle que l'élément $a_{ij}$ représente le nombre de liens du noeud $i$ vers le noeud $j$. Pour le réseau de Königsberg, on obtient ainsi (en notant les noeuds de la Fig. 1.1 en ordre croissant et de manière horaire en débutant par celui du haut) :

[1]Note historique : depuis la reconstruction de Königsberg (maintenant Kaliningrad) suivant la Seconde Guerre mondiale, deux ponts ont disparus et il est maintenant possible de traverser chaque pont restant qu'une seule fois ! Aux lecteurs intéressés de trouver quelles paires de ponts ont pu disparaître...

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 & 2 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 2 \\ 2 & 1 & 2 & 0 \end{pmatrix}. \tag{1.1}$$

**Definition 1.** Une *matrice d'adjacence* est une matrice carrée $A$ dont les éléments $a_{ij}$ sont donnés par la somme des poids des liens allant du noeud $i$ vers le noeud $j$.

Il y a deux éléments importants à remarquer dans la définition précédente. Premièrement, il est question des liens allant de *i vers j*, plutôt que des liens entre $i$ et $j$. Cette distinction est importante lorsqu'un réseau contient des liens ne pouvant être parcourus que dans une direction. Par exemple, si certains ponts de Königsberg étaient des sens uniques. On parle alors d'un *réseau dirigé* (voir Fig. 1.2(a)). Deuxièmement, la définition mentionne la somme des *poids* des liens et non pas simplement le nombre de liens. C'est pour tenir compte du fait que dans certains réseaux, deux liens ne sont pas nécessairement équivalents. Par exemple, dans un réseau social, un lien entre mari et femme aura souvent plus de poids qu'un lien entre deux collègues. Ou bien, pour en revenir au problème des ponts de Königsberg, supposons que certains ponts soient plus larges et permettent une circulation plus fluide que d'autres. On parlera maintenant d'un réseau *pondéré* (voir Fig. 1.2(b)). La question pourrait maintenant être une question d'optimisation : quel est le parcours qui permet de visiter les deux rives et les deux îles le plus rapidement possible ?



(a) (b)

FIG. *1.2:* **Réseaux dirigés et pondérés.** Le réseau des ponts de Königsberg en supposant que certains ponts sont à sens unique, résultant en un réseau dirigé en (a) ; et que certains ponts permettent une circulation plus fluide que d'autres, résultant en un réseau dirigé et pondéré en (b).

Pour ces nouvelles variations du réseau des ponts, la matrice d'adjacence prendrait la

forme :

$$\mathbf{A}_{\text{dirigé}} = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix} \qquad \mathbf{A}_{\text{pondéré}} = \begin{pmatrix} 0 & 3 & 0 & 2 \\ 3 & 0 & 3 & 1 \\ 0 & 3 & 0 & 1 \\ 1 & 1 & 2 & 0 \end{pmatrix} \tag{1.2}$$

où l'on a attribué les poids des liens selon l'épaisseur des traits de la Fig. 1.2(b).

Cette classification en différents types de réseaux est importante et illustre bien la généralité de cette forme de modélisation. En fait, pratiquement tout genre de systèmes peut être modélisé par un réseau si l'on choisit les bonnes propriétés. Par exemple, un problème de configuration d'horaire mensuel de médecins peut être solutionné en construisant un réseau possédant différents types de noeuds (multitypes) représentant les médecins à différentes instances temporelles et une seconde sorte de structure (bipartite) représentant les différents bureaux ; de multiples contraintes peuvent alors être forcées via les diverses propriétés du réseau. Certains iront même jusqu'à les utiliser pour solutionner des énigmes policières [32].

### 1.1.3 La physique statistique et les réseaux complexes

La représentation en matrice d'adjacence est une arme à double tranchant. Son avantage est d'offrir une connaissance explicite du réseau qui nous intéresse. Bien que ce soit une tâche ardue, il existe tout un arsenal d'outils pour s'attaquer à des problèmes à l'aide d'une matrice d'adjacence. Certains mathématiciens en ont d'ailleurs fait une spécialité [42]. L'autre côté de la médaille est qu'il devient rapidement très lourd de conserver explicitement une connaissance de tous les liens. Prenons par exemple la Fig. 1.3 qui présente la matrice d'adjacence du réseau d'interactions des personnages des Misérables de Victor Hugo, où beaucoup d'informations superflues sont gardées en mémoire, bien que l'oeuvre ne contienne que 77 personnages [40]. Considérant qu'il existe des données pour des réseaux allant jusqu'à plus de $10^7$ noeuds, il devient pratiquement impensable de ne travailler qu'à partir de la matrice d'adjacence. Et d'autre part, ce genre de description n'est pas très pratique lorsque nous n'avons pas une connaissance exacte de la structure ou que l'on cherche plutôt à avoir une description qualitative sans avoir à réinventer la roue à chaque fois qu'un lien s'ajoute.

Il n'est pas évident de partir d'une description instantanée de la position et de la vitesse de toutes les molécules d'un gaz quelconque dans un endroit quelconque et d'essayer d'en extraire les lois de la thermodynamique. Par contre, la physique statistique offre des outils bien adapté à ce genre d'analyse en supposant qu'on observe le système sous un moyennage à gros grains. C'est la philosophie que nous adapterons ici à la science des réseaux. C'est en fait avec cette philosophie que débute, en quelque sorte, la science des réseaux complexes telle que définie depuis une vingtaine d'années.



FIG. *1.3:* **Les Misérables.** Matrice d'adjacence des interactions entre les personnages du célèbre roman de Victor Hugo [40]. Les personnages sont numérotés selon leur ordre d'apparition et les éléments où il y a interaction sont en blanc, à l'exception des interactions correspondant à Valjean qui sont surlignées en rouge.

On se trouve alors confronté à la question qui nous accompagnera tout au long de notre travail de modélisation : quelle information doit-on conserver sur le réseau d'origine ? Ou plus directement : quelles sont les caractéristiques importantes des réseaux et quelles sont celles qui peuvent être inférées ?

### 1.1.4 Les réseaux Erdős–Rényi

L'extrême de la simplicité, c'est-à-dire la façon de garder le moins d'information possible sur le système, serait de simplement dire que le réseau des ponts est constitué de quatre noeuds et de sept liens. Sans plus d'information, il faudrait ensuite supposer que ces liens sont partagés de manière complètement aléatoire entre les noeuds. La Fig. 1.4 présente deux exemples de réseaux aléatoires construits à partir de cette logique et le vrai réseau de Königsberg avec une nouvelle disposition des noeuds. Est-il facile de distinguer le vrai du faux ?[2] Évidemment, certaines contraintes doivent être imposées aux réseaux aléatoires. Par exemple, un pont ne peut débuter et se terminer sur le même morceau de terre. On élimine ainsi les auto-liens (de l'anglais *self-loops*).

Ce type de réseaux aléatoires est communément appelé un réseau Erdős–Rényi, en l'honneur des mathématiciens Pál Erdős et Alfréd Rényi qui les premiers étudièrent en 1959 la connectivité de ces systèmes [17].

[2] Selon l'auteur : non.

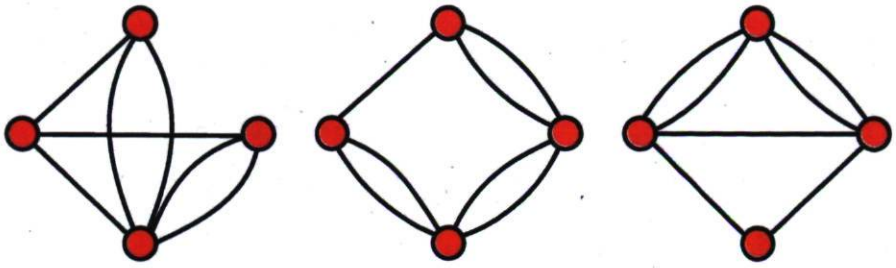


FIG. *1.4:* **L'urbaniste aléatoire.** Le premier de ces réseaux est en fait le réseau des ponts de Königsberg et les deux autres sont des réseaux aléatoires construits seulement à partir du nombre de noeuds et du nombre de liens.

**Definition 2.** Un réseau *Erdős–Rényi* (ER) est un réseau aléatoire $\Gamma$ construit de l'une de deux façons :

- $\Gamma(N, M)$ résulte de l'attachement aléatoire de $M$ liens sur $N$ noeuds ;
- $\Gamma(N, p)$ est construit en liant, avec probabilité $p$, chaque paire de noeuds d'un réseau de $N$ noeuds (aussi appelé *réseau de Poisson*).

Pourquoi est-ce que la *modélisation* du réseau des ponts de Königsberg en réseaux ER est si efficace ? Pour répondre à cette question, il est utile d'introduire le concept de séquence de degrés et de distribution en degré.

**Definition 3.** La *séquence de degrés* d'un réseau est une liste, $\{i_0, i_1, \ldots, i_{k_{max}}\}$, dont les éléments $i_k$ correspondent aux nombres de noeuds de degré $k$ dans le réseau pour $k \geq 0$. On peut donc en déduire le nombre de noeuds du réseau par $N = \sum_k i_k$ et le nombre de liens par $M = \sum_k k i_k/2$ puisque chaque lien est constitué de deux degrés.

**Definition 4.** La *distribution en degré* d'un réseau est la distribution $\{p_0, p_1, \ldots, p_{k_{max}}\}$, où $p_k$ est la proportion des noeuds du réseau ayant un degré $k$. Elle se doit évidemment d'être normalisée, $\sum_k p_k = 1$ et on peut en retirer facilement le degré moyen du réseau $\langle k \rangle = \sum_k k p_k$.

Par exemple, la séquence de degrés du réseau des ponts de Königsberg correspond à {0, 0, 0, 3, 0, 1} pour un total de 4 noeuds, 14 degrés et donc 7 liens ; et la distribution en degré correspondante est {0, 0, 0, 3/4, 0, 1/4} pour un degré moyen de 3.5. Pour un réseau ER $\Gamma(N, M)$, on aura une simple distribution binomiale correspondant à la probabilité d'attribuer $k$ des $2M$ degrés disponibles au même noeud :

$$p_k = \binom{2M}{k} (1/N)^k \, (1 - 1/N)^{2M-k} \tag{1.3}$$

La Fig. 1.5 compare la distribution en degré du réseau des ponts avec la distribution en degré *moyenne* d'un ER $\Gamma(4, 7)$ telle que dictée par l'Éq. (1.3). Les séquences sont visiblement

similaires dans le sens où elles sont toutes deux concentrées autour de la même moyenne, mais ce n'est pas suffisant pour bien modéliser un réseau. En fait, la densité du réseau des ponts facilite également sa modélisation. Puisqu'en moyenne un noeud a plus de $N-1$ liens, il n'est pas surprenant de voir apparaître des *triangles* et des *double-liens* dans les ER comme dans le vrai réseau.

**Definition 5.** La *densité* d'un réseau, ou de l'un de ses sous-ensembles, est définie par la fraction des liens qui existent par rapport à ceux pouvant exister $n(n-1)/2$, où $n$ est le nombre de noeuds dans l'ensemble considéré et où l'on suppose qu'une paire de noeuds ne peut être liée qu'une seule fois.

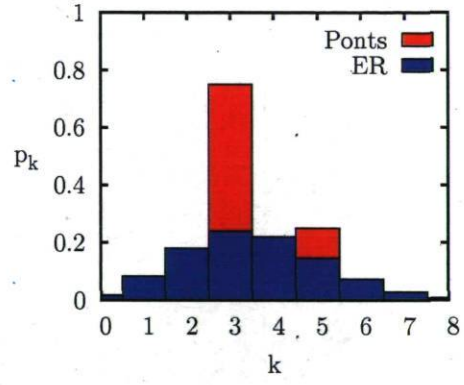



FIG. *1.5:* **Distribution en degré du réseau de Königsberg.** Comparaison de la distribution en degré du réseau des ponts de Königsberg et d'une construction aléatoire en réseau Erdős–Rényi.

Bref, la similarité des réseaux de la Fig. 1.4 est principalement due à l'homogénéité de la séquence de degrés du réseau et à sa densité qui force des structures particulières. Dans les deux sections suivantes, nous nous attarderons à ces deux aspects, dans cet ordre, dans le but de tester la généralité (ou universalité) des réseaux ER comme outil de modélisation. Ainsi, la première étape est de chercher un comportement universel (i.e. commun à des réseaux de diverses natures) dans la distribution en degré de réseaux réels.

## 1.2 Réseaux libres d'échelle

Dans les prochaines sections, nous verrons essentiellement en quoi les réseaux réels diffèrent des réseaux ER. Pour ce faire, on se tournera vers différentes découvertes fondamentales des douze dernières années.

### 1.2.1 Barabási et Albert, 1999

Peu avant le tournant du millénaire, l'étude de systèmes complexes vécut un chamboulement d'importance : l'arrivée de bases de données décrivant de larges systèmes. Avec la possibilité de s'attaquer à des réseaux allant jusqu'au million de noeuds, une vague de physiciens (principalement issus de la physique statistique) se lança dans l'aventure. Suivant la logique que nous avons exposée jusqu'ici, une des premières questions posées concerna la distribution en degré de ces réseaux.

Une étude sur le sujet fut publiée dans le journal *Science* en 1999 par Albert-László Barabási et Réka Albert [9]. Dans leur article, ils mirent en évidence une propriété surprenante dans la distribution en degré de trois réseaux de natures différentes (voir Fig. 1.6).

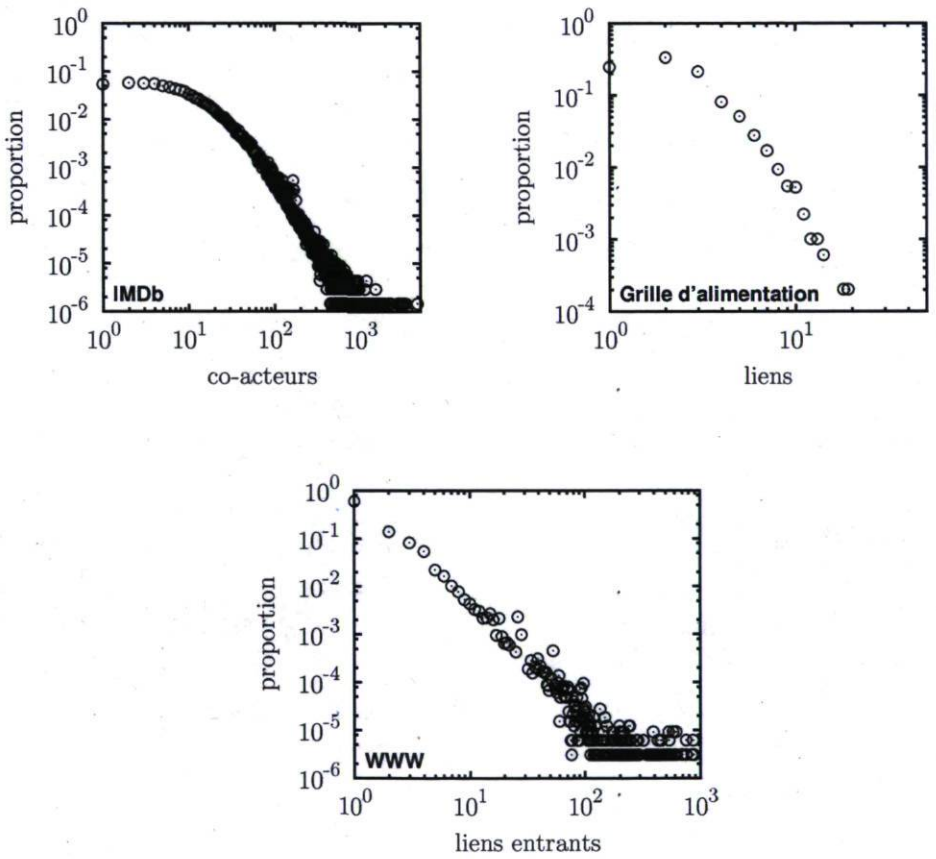



FIG. *1.6:* **Réseaux libre d'échelle.** Les résultats de Barabási et Albert [9] sur les collaborations entre acteurs telles que détaillées dans l'Internet Movie Database ($\sim 10^6$ noeuds), le réseau d'alimentation électrique des états de l'ouest américain ($\sim 10^3$ noeuds) et les sites web de l'Université Notre-Dame ($\sim 10^5$ noeuds).

À la Fig. 1.5, on voyait que le réseau des ponts de Königsberg était similaire aux réseaux ER parce que sa distribution en degré était fortement homogène, les noeuds ayant tous un degré de l'ordre du degré moyen. Évidemment, ce n'est pas surprenant considérant que l'on avait alors un système ne possédant que 4 noeuds ! Or, les résultats de Barabási et Albert montrent non seulement que les trois réseaux considérés ont des distributions fortement inhomogènes, mais que tous trois possèdent le même comportement de queue lourde (*fat tail*). On entend ici que certains éléments se retrouvent avec un degré de plusieurs ordres de grandeur supérieur au degré moyen du système.

En fait, les queues des distributions suivent toutes des lois de puissance $p_k \propto k^{-\gamma}$. Ces lois sont dites *indépendantes d'échelle* parce qu'elles conservent leur forme sous dilatation, i.e. :

$$p_{\lambda k} = \lambda^{-\gamma} p_k \,. \tag{1.4}$$

Pour une distribution en degré, cela veut dire que le ratio entre $p_{10}$ et $p_{100}$ est potentiellement le même qu'entre $p_{100}$ et $p_{1000}$ ! Ces systèmes sont donc indépendants d'échelle puisqu'il est impossible d'y identifier un degré (ou une échelle) caractéristique. Cette propriété est facile à observer sur les résultats de la Fig. 1.6, puisqu'elle implique que les distributions possèdent des queues linéaires sur un graphique log-log (comme une loi de puissance, la pente de cette droite correspond ici à l'exposant d'échelle $\gamma$ de la distribution). En ce sens, les réseaux sont similaires à bien d'autres systèmes complexes qui présentent le même genre de comportement d'échelle :

- distribution des genres biologiques par leur nombre d'espèces [82] ;
- distributions de mots par leur nombre d'apparitions dans des extraits de prose [84] ;
- distributions de salaires [22] ;
- tailles des cratères lunaires et intensités des tremblements de terre [54] ;
- activités de recherche par université [67] ;
- expression de gènes [20] ;
- et pratiquement tout genre d'activité humaine [84; 70; 54].

### 1.2.2 Conséquences de l'indépendance d'échelle

En considérant la liste de systèmes libres d'échelle ci-dessus[3], on est en droit de se demander qu'est-ce qu'il y a de si important au fait que les réseaux complexes présentent une telle organisation.

Pour illustrer les conséquences de cette découverte, considérons une dynamique simple : la robustesse face à une attaque contre le réseau. Le but de ce problème est de déterminer à

[3] Qui, soit dit en passant, ne représente que le sommet de l'iceberg.

quel point la structure d'un réseau est affectée par le retrait ciblé d'une fraction $\epsilon$ des noeuds, ce qui simule l'efficacité d'une campagne de vaccination ou la résistance d'un réseau informatique aux attaques de pirates. Pour quantifier l'effet désiré, on empruntera un concept de la théorie de la percolation [10].

**Definition 6.** Les *composantes* d'un réseau sont les plus grands ensembles de noeuds tels que tous les noeuds d'une même composante sont rejoignables entre eux en suivant un nombre quelconque de liens.

**Definition 7.** De façon pratique, la *composante géante* d'un réseau est simplement la plus grande composante. Théoriquement, c'est la seule composante *extensive* du réseau, ce qui implique qu'elle occupe une fraction macroscopique $S$ du système même dans la limite infinie.

Dans le problème de robustesse contre les attaques, on cherche donc à savoir à quel point la composante géante du réseau est robuste face aux retraits ciblés de noeuds. Pour prendre un cas plus parlant, prenons la structure l'Internet au niveau des systèmes autonomes (SA) [4] circa 2006 (voir Fig. 1.7(a)). Le retrait de noeuds correspond ici à une attaque contre les SA de plus haute importance, donc de plus haut degré, et l'on cherche à voir si l'Internet possède suffisamment de voies de routage secondaires pour être robuste. Pour tester cette propriété, on attaque directement la portion $\epsilon$ des noeuds de plus hauts degrés.

Les résultats de cet exercice sont présentés à la Fig. 1.7(b) et illustrent bien la vulnérabilité des réseaux libres d'échelle face aux attaques ciblées. C'est que les noeuds de plus haut degré dans les réseaux libres d'échelle (typiquement nommés les *hubs*) ont un nombre si élevé de liens que le réseau se défait complètement si on les retire. Les hubs sont cependant complètement absents du réseau équivalent ER (réseau aléatoire avec autant de noeuds et de liens), ce qui explique sa plus grande robustesse. Quoique catastrophique dans le cas simulé (attaques sur Internet), cette propriété peut s'avérér être une bénédiction lorsque l'on connaît les noeuds de haut degré et que l'on cherche à contrôler le réseau (e.g. limiter la propagation d'un polluant dans une toile alimentaire).

Par opposition à cette vulnérabilité face aux attaques ciblées, on peut également étudier la résistance de l'Internet face aux pannes aléatoires de SA. Cette fois, on désactive une fraction $\epsilon$ des noeuds du réseau de façon aléatoire, indépendamment de leur degré. On présente les résultats de cet exemple sur la Fig. 1.7(c). Ce que l'on peut rapidement constater est que la chute de la composante géante est pratiquement aussi lente pour l'Internet que pour son équivalent ER. Ce qui veut dire que le réseau réel est robuste face à ce type de défaillances malgré sa vulnérabilité aux attaques ciblées.

[4] Les ensembles de réseaux informatiques avec une procédure de routage interne prédéterminée.

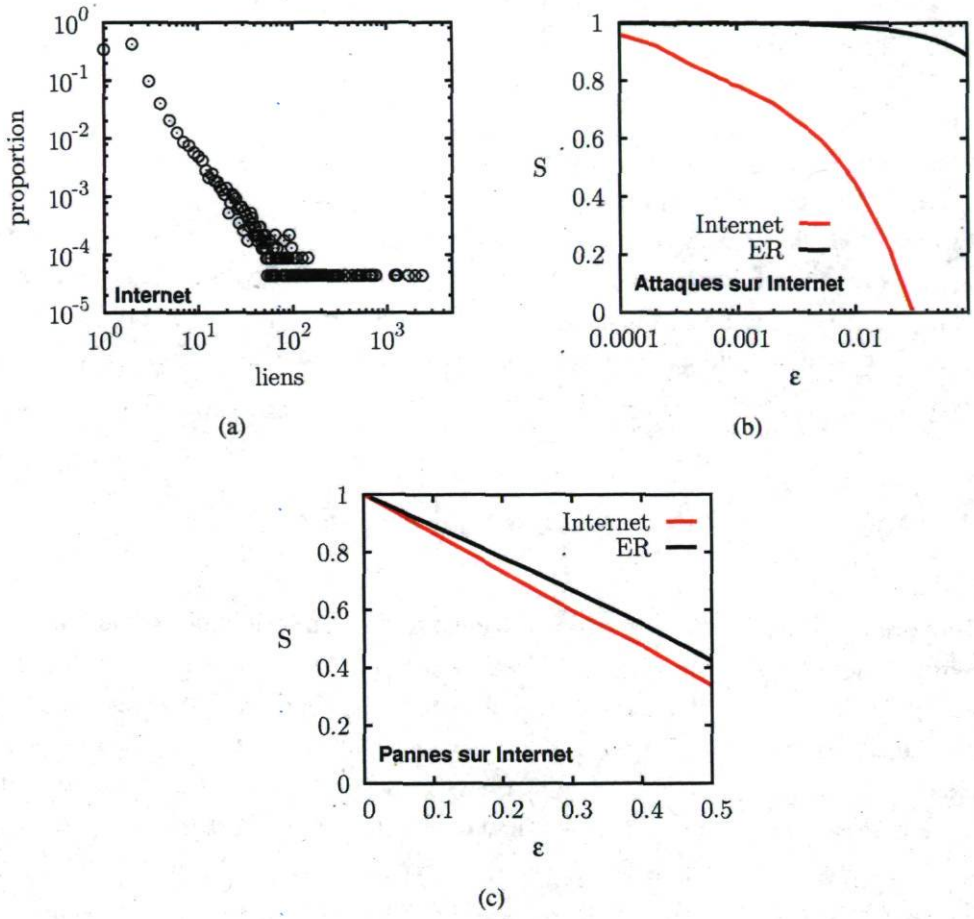



FIG. *1.7:* **Pannes et attaque sur l'Internet** (a) La distribution en degré de l'Internet. (b) La perte de connectivité (décroissance de la composante géante $S$) sous attaques ciblées vers les noeuds de haut degré. (c) Même chose sous pannes aléatoires de noeuds.

Ce résultat est assez simple à expliquer considérant la distribution en degré de la Fig. 1.7(a). Sa forte hétérogénéité et son asymétrie impliquent que si l'on retire un noeud au hasard, on est beaucoup plus probable de retirer un noeud de bas degré (sous la moyenne) qu'un des *hubs* du réseau. Pour s'en convaincre, remarquons que plus de 75% des noeuds sont de degré un ou deux, donc sous le 4.2 de moyenne. Il faut donc éplucher beaucoup de noeuds sans importance avant d'avoir une bonne probabilité de toucher les noeuds de haut degré[5]. Le réseau ER présente le même genre de robustesse peu importe de quelle façon on lui retire ses noeuds, puisque tous les noeuds sont plus ou moins équivalents.

[5]Voilà pourquoi il est difficile de protéger un réseau informatique contre un virus : les virus se retrouvent préférentiellement dans les hubs, alors que les efforts de protection sont plus aléatoires puisqu'il est difficile de connaître, par exemple, le nombre de contacts électroniques entre différentes stations de travail. Il est donc utile de tenter de propager la protection par les mêmes voies que l'infection [45].

Cette dualité entre robustesse et vulnérabilité des réseaux libres d'échelle a été explorée pour la première fois par Albert *et al.* en 2000 [3]. Il existe cependant bien d'autres propriétés inhérentes aux réseaux libres d'échelle. L'une d'entre elles est particulièrement d'intérêt pour notre projet. On verra au Chap. 4 que la capacité d'une maladie à se propager sur un réseau dépend principalement du second moment de la distribution en degré (gouvernant le nombre de liens que l'on peut suivre après être arrivé sur un noeud en suivant un lien aléatoire : le *degré sortant*). Or, il est bien connu que les moments d'une distribution en loi de puissance d'exposant $\gamma$ explosent de sorte que les moments $m \geq \gamma - 1$ divergent. Par conséquent, le second moment est typiquement beaucoup plus élevé que le degré moyen, voir infini, rendant les réseaux libres d'échelle extrêmement vulnérables aux épidémies. Plus de détails seront donnés au Chap. 4.

### 1.2.3 Modèle canonique de réseaux aléatoires

L'importance de la distribution en degré vient d'être illustrée à l'aide de quelques exemples. Malheureusement, le seul modèle de réseaux aléatoires que l'on a présenté jusqu'ici, les réseaux ER, ne nous accorde aucune liberté quant à la forme de la distribution en degré. On présente donc ici un modèle plus général de réseaux aléatoires, qui au lieu de fixer le nombre de noeuds et le nombre de liens, requiert plutôt le nombre de noeuds et la distribution en degré. Ce nouveau modèle est plus clairement décrit en tant qu'algorithme :

i. générer une séquence de degrés $\{d_i\}$ de longueur $N$ aux valeurs soumises à la distribution en degré $\{p_k\}$ ;

ii. pour chaque $i$, produire $d_i$ bouts de liens identifiés $i$ ;

iii. assigner en paires tous les bouts de liens de façon aléatoire ;

iv. si désiré, tester le réseau résultant pour des doubles liens (paires identiques) ou pour des auto-liens (paires $\{i, i\}$) ;

v. si le test est positif, refaire la séquence du début jusqu'à l'obtention d'un réseau satisfaisant toutes les contraintes.

On parlera donc d'un modèle *canonique* de réseaux aléatoires. L'analogie étant que le modèle ER fixe le nombre de liens (énergie fixe) décrivant ainsi le système de façon *micro-canonique*, alors que ce nouveau modèle ne fixe que le degré moyen (énergie moyenne). En anglais, on parle plutôt du *configuration model*, car, dans son ensemble canonique, il explore toutes les configurations possibles de réseaux avec la même distribution en degré.

## 1.3 Réseaux modulaires et les *small-worlds*

Maintenant que nous avons développé une certaine intuition quant à la façon dont les liens peuvent être distribués à l'intérieur de réseaux complexes, nous devons prendre en compte un peu plus d'information. Pour ce faire, on adopte un point de vue un peu moins *local* autour des noeuds pour se poser différentes questions. Comment ces liens sont-ils organisés ? Vers qui mènent-ils ? Existent-ils des corrélations entre différents liens ?

Précisons d'abord que notre intuition sur ce que devrait être un réseau est en contradiction flagrante avec nos modèles de réseaux aléatoires. De notre point de vue, dans les différents réseaux qui nous entourent (électriques, d'information, mais principalement les réseaux sociaux), il semble possible de définir un espace (non-euclidien) et une notion de proximité entre les noeuds de sorte que *le voisin de mon voisin est mon voisin*. Ce genre de corrélations à courte portée n'est cependant pas pris en compte ni dans le modèle ER ni dans le modèle canonique de réseaux aléatoires. Pourtant, ce n'est pas un incroyable hasard que je connaisse mes quatre collègues du groupe de recherche en réseaux complexes, et qu'ils se connaissent également les uns les autres. Ces structures, ici une clique de cinq noeuds[6], sont inhérentes aux réseaux que nous formons, que nous construisons et qui nous constituent.

En comparant le réseau des ponts de Königsberg à des réseaux ER équivalents, on a mentionné que la densité du système réel favorisait sa modélisation en réseau aléatoire. C'est que le réseau des ponts était constitué de multiples triangles, et ce genre de structure ne peut émerger de manière significative dans un modèle aléatoire que si chaque paire de noeuds a une probabilité importante d'être liée. Et comme typiquement le degré moyen est une propriété *intensive* des réseaux complexes[7], la probabilité d'émergence de ces structures tend vers zéro lorsque le réseau grossit.

Pour incorporer cette *structure sociale* dans nos modèles, il faut d'abord tenter de mieux quantifier son importance. C'est ce qui a été fait dans une autre étude célèbre de la science des réseaux, par Duncan Watts et Steven Strogatz dans un article publié dans le journal *Nature* en 1998 [80].

### 1.3.1 Watts et Strogatz, 1998

De prime abord, notre intuition nous dit que les réseaux sociaux sont *agrégés* en groupes densément connectés où *l'ami de mon ami est mon ami*. Cette impression entre en fort

[6] Sans offense.

[7] Les gens n'auraient pas deux fois plus d'amis si la Terre était deux fois plus peuplée.

contraste avec certains résultats empiriques telle qu'une expérience menée par Stanley Milgram en 1967 [77]. Dans son expérience, Milgram envoie des paquets postaux aléatoirement à des citoyens américains leur demandant de passer le paquet à un individu à Boston. Le paquet contient par conséquent des informations sommaires sur l'individu cible. Si les individus choisis connaissent la cible, ils doivent lui transférer directement le paquet ; sinon, ils doivent passer le paquet à l'une de leurs connaissances qui selon eux, serait plus probable de connaître la cible. Selon les paquets qui se sont éventuellement rendus à destination, Milgram conclut que deux individus choisis au hasard dans les États-Unis sont en moyenne à trois degrés de séparation l'un de l'autre. On dit alors que le réseau social américain possède une distance moyenne de trois. Au niveau planétaire, la légende urbaine veut que la distance moyenne soit de six[8].

**Definition 8.** La *distance* entre deux noeuds $\ell_{ij}$ est définie par le nombre de liens composant le *plus court chemin* reliant ces deux noeuds $i$ et $j$. La distance moyenne $\ell$ est ainsi simplement donnée par :

$$\ell = \frac{1}{\binom{N}{2}} \sum_{i,j} \ell_{ij} \,. \tag{1.5}$$

Le paradoxe soulevé par Watts et Strogatz est que si nos interactions sociales sont typiquement à courte portée et dans des groupes denses, il devrait être beaucoup plus difficile de se rendre à une personne éloignée (socialement ou géographiquement) que ne le démontre l'expérience de Milgram. Pour illustrer le paradoxe, imaginez vous assis dans un large stade (ou dans le réseau de la Fig. 1.8(a)) et imaginez devoir passer un message à la personne assise complètement à l'opposé. Vous devriez alors passer le message à votre voisin, qui le passera à son voisin et ainsi de suite jusqu'à se rendre à la cible. Or, l'expérience nous dit que les distances à parcourir sont probablement beaucoup plus courtes, une propriété typiquement associée aux réseaux aléatoires (voir Fig. 1.8(b)).

Pour s'en convaincre, on peut caractériser le comportement de la distance moyenne d'un réseau de taille $N$ en fonction de sa structure. Par exemple, pour le cas du réseau de la Fig. 1.8(a) (i.e. le stade de l'exemple précédent), il est facile de voir que la distance moyenne aura un comportement linéaire en $N$ puisque le réseau est essentiellement une chaîne de voisinages : $\ell \propto N$. Pour un réseau ER suffisamment grand sans être trop dense, i.e. $\langle k \rangle \ll N$, on aura plutôt une structure en arbre.

**Definition 9.** Une *structure en arbre* est une structure telle que chaque lien suivi mène à un nouveau noeud. On peut donc monter ou descendre dans le réseau, mais on ne peut jamais revenir sur ses pas via une *boucle*.

Ainsi, chaque nouveau noeud rejoint nous donne accès à $\langle k \rangle$ nouveaux noeuds[9]. On peut

[8] L'expression *Six degrés de séparation* est probablement due à John Guare, auteur d'une pièce de ce nom.
[9] Fait intéressant, le degré sortant moyen d'un large réseau ER est égal au degré moyen.

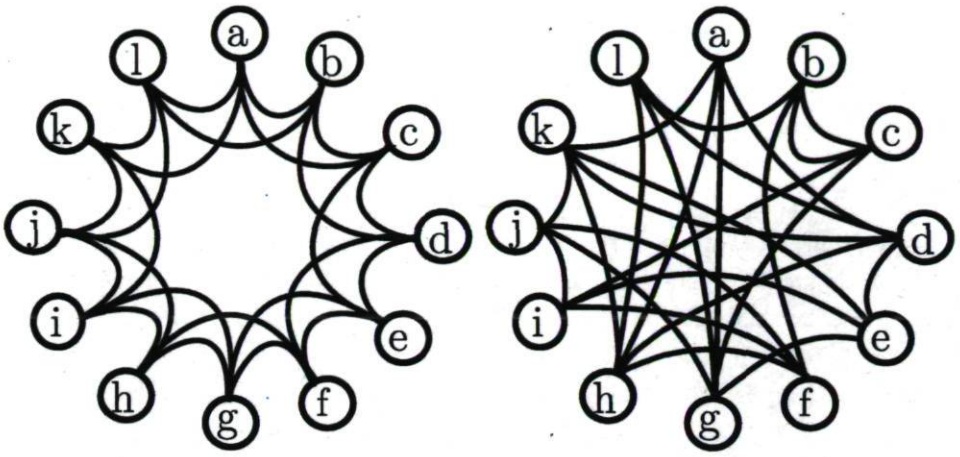



FIG. *1.8:* **Réseaux ordonnés et désordonnés** (a) Un réseau ordonné : les connexions se font selon la proximité des noeuds. (b) Le même réseau avec chaque lien aléatoirement reconnecté.

en conclure que le nombre $n$ de noeuds accessibles en $\ell$ pas est de $n \propto \langle k \rangle^{\ell}$ ou inverser cette relation et obtenir que la distance moyenne d'un réseau d'ER croît logarithmiquement avec la taille du réseau : $\ell \propto \log N$. Ce qui implique que la distance moyenne est toujours plus petite dans un réseau aléatoire que dans un réseau basé sur la proximité des noeuds.

Pour vérifier pour quelle raison notre intuition nous fait défaut, Watts et Strogatz définissent une seconde quantité, le coefficient d'agrégation.

**Definition 10.** Le *coefficient d'agrégation* d'un réseau mesure à quel point il est probable que *l'ami de mon ami soit mon ami*. On le définit par le ratio du nombre de triangles $N_{\Delta}$ et du nombre de triades $N_{\vee}$ (paires de liens partageant un noeud) multiplié par trois :

$$C = \frac{3N_{\Delta}}{N_{\vee}}, \tag{1.6}$$

où le facteur 3 tient compte du fait qu'un triangle contribue à trois triades. On a ainsi une quantité variant de zéro à un pour évaluer la présence de triangles dans un réseau.

Ainsi, $C$ et $\ell$ devraient être grands dans un réseau ordonné par proximité, alors qu'ils seront tous deux petits dans un réseau aléatoire. Watts et Strogatz peuvent ainsi vérifier si les réseaux réels ressemblent plus à la Fig. 1.8(a) ou à la Fig. 1.8(b) en mesurant $C$ et $\ell$ chez de vrais réseaux. Les résultats obtenus dans leur étude sont présentés au Tableau 1.1.

Avec un grand coefficient d'agrégation et une petite distance moyenne, les réseaux réels semblent se situer à mi-chemin entre les réseaux ordonnés et les réseaux désordonnés. Watts

| réseau | $\ell_{\text{vrai}}$ | $\ell_{\text{ER}}$ | $C_{\text{vrai}}$ | $C_{\text{ER}}$ |
|---|---|---|---|---|
| IMDb | 3.65 | 2.99 | 0.79 | 0.00027 |
| Électrique | 18.7 | 12.4 | 0.080 | 0.005 |
| *C. elegans* | 3.65 | 2.99 | 0.28 | 0.05 |

TAB. *1.1:* **Résultats de Watts et Strogatz.** Distance moyenne $\ell$ et coefficient d'agrégation $C$ pour différents réseaux : collaborations entre acteurs de l'IMDb, grille d'alimentation électrique de l'ouest américain et réseaux neuronaux du ver *C. elegans*. Pour chaque système, les valeurs sont comparées à celles obtenues sur un réseau possédant le même nombre de noeuds et le même degré moyen, mais aux connexions aléatoires : un réseau ER.

et Strogatz réfèrent à cette structure comme étant un *small world*, car il s'agit bel et bien d'un monde où certaines règles de proximité s'appliquent, mais des liens longues distances lui procurent une structure aussi navigable que celle d'un réseau aléatoire (voir Fig. 1.9).

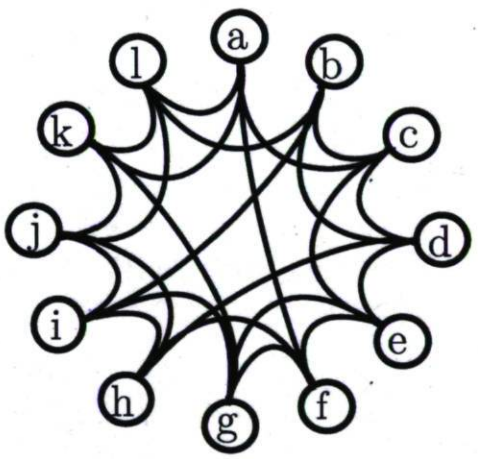



FIG. *1.9:* **Réseau *small-world* : entre l'ordre et le désordre** Les réseaux réels semblent partager des propriétés des systèmes ordonnés (agrégation) et des systèmes aléatoires (petite distance moyenne).

### 1.3.2 La structure communautaire

Une critique que l'on peut faire à l'analyse de Watts et Strogatz est la quantification d'*ordre* ou d'agrégation qu'ils ont choisie. Considérons par exemple les trois motifs de réseaux illustrés à la Fig. 1.10. Il est difficile de se faire une bonne idée de l'organisation des liens à partir du coefficient d'agrégation. En fait, un peu comme la distribution en degré est la meilleure façon de décrire la connectivité d'un réseau (plutôt que simplement le degré moyen), on se devra d'introduire des distributions en *structure communautaire* pour mieux définir les topologies complexes entre l'ordre et le désordre.

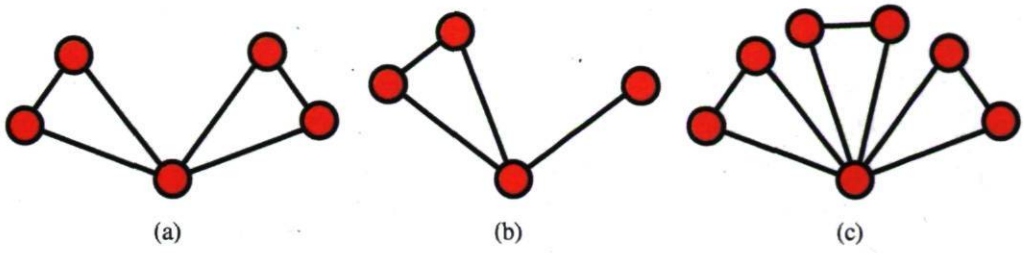



FIG. *1.10:* **Motifs de réseaux plus ou moins ordonnés.** Il est loin d'être clair quels motifs sont plus ordonnés, agrégés ou au contraire aléatoires, et leur coefficient d'agrégation ne sont pas nécessairement plus révélateurs : (a) $C = 3/5$ ; (b) $C = 3/5$ et (c) $C = 3/7$.

La structure communautaire est définie par le regroupement des noeuds en groupes (*communautés*) plus densément connectés que la moyenne du réseau (voir Fig. 1.11). Pour la spécifier, nous emploierons deux distributions.

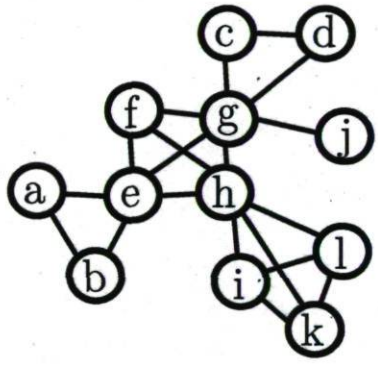



FIG. *1.11:* **Exemple de réseau à structure communautaire.** La grande majorité des vrais réseaux ont une topologie à structure communautaire comme celle illustrée ici plutôt que celle du modèle de Watts et Strogatz de la Fig. 1.9.

**Definition 11.** La *distribution en taille* des communautés est l'ensemble des proportions $p_n$ (normalisées) de structures composées de $n$ noeuds du réseau.

**Definition 12.** La *distribution en appartenance* des noeuds est l'ensemble des proportions $p_m$ (normalisées) de noeuds appartenant à $m$ communautés.

Par exemple, sur l'exemple de la Fig. 1.11, nous avons une communauté de taille 2 ($\{g, j\}$), deux communautés de taille trois ($\{a, b, e\}$ et $\{c, d, g\}$) et deux communautés de taille quatre ($\{e, f, g, h\}$ et $\{h, i, k, l\}$). De même, nous avons 9 noeuds appartenant à une seule communauté ($a$, $b$, $c$, $d$, $f$, $i$, $j$, $k$ et $l$), deux noeuds appartenant à deux communautés ($e$ et $h$) et un seul noeud appartenant à trois communautés ($g$). À partir de ces listes, il est aisé de calculer les distributions en taille et en appartenance.

Lorsqu'est venu le temps de calculer une distribution en degré, il a été assez direct de calculer le nombre de liens par noeud à partir des bases de données. Pour les deux distributions décrivant la structure communautaire, on doit d'abord définir ce qu'est une communauté et ensuite identifier toutes celles existant dans un réseau donné. Comme nous le verrons sommairement dans cette section en survolant deux méthodes de détection de structure communautaire, il s'agit là d'une tâche extrêmement complexe. Supposons que l'on retire le lien entre les noeuds $e$ et $g$ dans la Fig. 1.11, définissons-nous encore le groupe $\{e, f, g, h\}$ comme une communauté ?

### Communautés de noeuds

Une des premières tentatives de détection de communautés à avoir porté fruits est la *méthode par percolation de cliques* [62]. Dans cet algorithme, on définie les communautés à partir des cliques qu'elles contiennent.

**Definition 13.** Une *clique-k* est un ensemble de $k$ noeuds *complètement* connectés.

**Definition 14.** Une *communauté-k* est une communauté de noeuds définie par un ensemble de *cliques-k* tel que chaque clique partage $k-1$ noeuds avec au moins une autre clique de l'ensemble.

Les algorithmes de *percolation de cliques* identifient d'abord toutes les *cliques-k* pour ensuite les regrouper en chaîne lorsque deux se chevauchent par $k-1$ noeuds. Dans cette méthode, il est primordial de bien comprendre le rôle que joue le paramètre $k$. Si l'on utilise un $k$ trop élevé, on cherche à utiliser de trop grosses cliques comme base de la percolation, il est donc bien probable de ne pas en identifier assez pour être en mesure de détecter les communautés (ou on ne détectera que les communautés grandes et denses). De même, avec un $k$ trop petit, on risque d'avoir trop de petites cliques interconnectées ce qui a typiquement comme conséquences de surestimer la taille des communautés. Par exemple, si on prend $k = 2$, les communautés détectées correspondront aux composantes du réseau, ce qui nous informe très peu sur la structure. Ainsi, le paramètre $k$ joue un peu le rôle d'un focus, contrôlant le niveau d'organisation du réseau que l'on désire observer. La validation de ce genre d'algorithmes est typiquement fait sur des réseaux ayant une structure communautaire connue (voir Fig. 1.12).

### Communautés de liens

Une approche un peu plus récente a également su démontrer son efficacité pour détecter les communautés d'un réseau. Cette approche est basée sur un point de vue différent, où les liens, plutôt que les noeuds, sont assignés à des communautés. Le but est d'identifier des

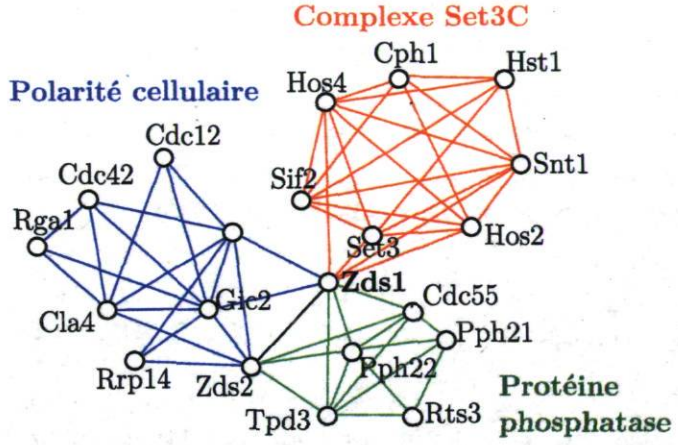



FIG. *1.12:* **Détection des complexes de protéines de la levure *S. Cerevisiae*.** Les communautés détectées dans un réseau d'interactions de protéines peuvent ensuite être identifiées en tant que complexes. Ici, une percolation de cliques avec $k = 4$ est utilisée pour détecter les trois complexes autour de la protéine Zds1.

interactions qui ont lieu dans le même environnement, donc des liens reliant des voisinages similaires. Comme mesure de similarité, on emploie habituellement l'indice de Jaccard.

**Definition 15.** Le *voisinage* d'un noeud est définie comme l'ensemble des voisins du noeud ; la taille d'un voisinage est donc égale au degré du noeud d'intérêt. On peut ensuite parler de *second voisinage*, en s'intéressant aux voisins des éléments du premier voisinage, et ainsi de suite.

**Definition 16.** L'*indice de Jaccard* est une mesure de similarité entre deux ensembles $A$ et $B$ définie par :

$$J_{AB} = \frac{|A \cap B|}{|A \cup B|} . \tag{1.7}$$

On peut ensuite procéder comme suit pour identifier les communautés de liens [1] :

i. pour tout noeud $i$, identifier toutes les paires de liens $\{e_{ij}, e_{ik}\}$ émergeant de $i$ ;

ii. pour chaque paire $\{e_{ij}, e_{ik}\}$, identifier les voisinages $A$ et $B$ des noeuds $j$ et $k$ se trouvant aux extrémités des liens ;

iii. calculer l'indice de Jaccard entre les voisinages $A$ et $B$ ;

iv. les liens d'une paire ayant un coefficient de Jaccard supérieur à un seuil donné $J_c$ sont assignés à la même communauté de liens ;

v. les noeuds d'une communauté sont ensuite déterminés en allant voir les extrémités de chaque lien qui lui est assigné.

Les avantages de cette méthode sont nombreux. Entre autres, la possibilité d'identifier des communautés emboîtées (*nested communities*) où tous les noeuds d'une petite communauté appartiennent également à une plus grande communauté moins dense. De plus, l'algorithme ne possède pas de limite de résolution puisque les liens n'ayant pas été assignés à une communauté constituent à la fin des communautés de taille deux (versus la percolation de cliques qui ne peut détecter les communautés de taille inférieure à $k$).

## 1.4 L'auto-similarité des réseaux complexes

À la section précédente, notre but était de prendre un point de vue moins local et plus éloigné autour des noeuds pour voir l'organisation des liens en terme de regroupement de noeuds. Nous ferons ici la même chose. Reculons encore un peu plus notre perspective sur le réseau et tentons de déterminer comment ces groupes (ou communautés) sont eux-mêmes interconnectés.

### 1.4.1 Song *et al.*, 2005

La lettre publiée par Song *et al.* dans *Nature* en 2005 [74] nous renseigne beaucoup sur l'organisation des réseaux complexes au-delà du lien.

De manière simplifiée, puisque l'indépendance d'échelle est typiquement associée aux géométries fractales, leur idée est de vérifier si les réseaux complexes n'auraient pas une structure auto-similaire. Pour ce faire, ils développent une simple *décomposition en boîtes* où des noeuds sont placés dans une même boîte s'ils sont tous séparés par une distance inférieure à un $\ell_b$ donné. On s'intéresse ensuite à savoir comment ces boîtes ou groupes de noeuds sont interconnectés (deux boîtes sont connectées s'il existe au moins un lien entre leurs noeuds respectifs).

La longueur $\ell_b$ contrôle la grosseur des boîtes. On peut ainsi décomposer le réseau à différents niveaux de résolution pour ensuite s'intéresser à caractériser comment les boîtes sont interconnectées via les liens restants.

Le résultat intéressant de cette étude est que les distributions en degré de certains réseaux sont auto-similaires sous cette décomposition en boîtes. C'est à dire que les distributions en degré des boîtes de taille $\ell_b$ sont presque identiques à la distribution d'origine, indépendamment de $\ell_b$. Puisque cette décomposition en boîtes est similaire à une méthode grossière de détection de communautés, on peut plutôt s'intéresser à comparer la distribution en degré des noeuds et la distribution en degré communautaire des communautés.

**Definition 17.** Le *degré communautaire* d'une communauté représente le nombre de liens qu'elle possède avec d'autres communautés. Typiquement, ces liens correspondent au fait que deux communautés partagent au moins un noeud tel que le degré communautaire correspond au nombre de communautés qui chevauchent la communauté d'intérêt.

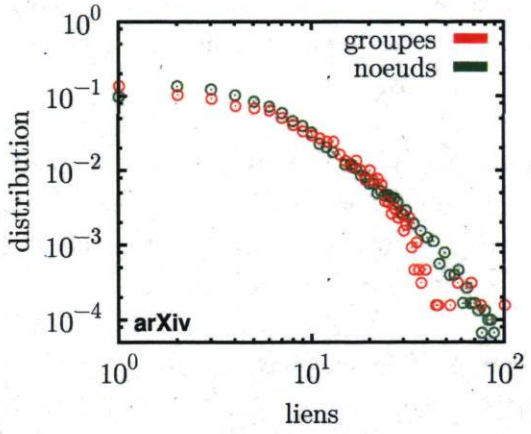



FIG. *1.13:* **Auto-similarité des niveaux d'organisation du *cond-mat arXiv*.** Les distributions en degrés entre les auteurs de l'*arXiv* et des différentes communautés détectées par l'algorithme de communautés de liens.

À la Fig. 1.13, on présente le résultat de cette comparaison pour le réseau de collaborations entre auteurs dans l'archive de matière condensée *cond-mat arXiv* circa 2005. Dans ce réseau, les noeuds sont des auteurs scientifiques et ils sont liés s'ils ont écrit au moins un article ensemble (plus de détails à la Sec. 2.10). On peut y voir le même genre d'auto-similarité telle qu'observée par Song *et al.* sur le World-Wide Web et des réseaux d'interactions de protéines. Cependant, l'auto-similarité entre l'organisation des communautés et l'organisation des liens (c'est-à-dire sans faire intervenir une décomposition en boîtes *ad hoc*) n'avait jamais été observée aussi directement avant nos travaux [29].

## 1.5 Propriétés universelles des réseaux complexes

Le but de ce premier chapitre était non seulement de mettre en lumière certaines propriétés d'organisation, mais également de mettre l'emphase sur la façon dont les réseaux réels diffèrent généralement des réseaux aléatoires. Ultimement, l'idée est de mettre le doigt sur ce qui doit être incorporé dans les modèles pour avoir une description à la fois générale et réaliste des réseaux complexes.

### 1.5.1 Vérification et application : l'épidémiologie sur réseaux

Lorsqu'un modèle est développé, il est simple et direct de vérifier sa qualité en essayant de reproduire la topologie d'un système réel via quelques propriétés d'organisation. Cependant, nous sommes toujours limités à ce niveau par ce que nous sommes en mesure de quantifier. Il est évident qu'il existe bien d'autres niveaux d'organisation que ceux que nous avons décrits dans ce chapitre. Si nous ne les connaissons pas, comment pouvons nous vérifier leur importance ? Comment tester notre aptitude à reproduire la topologie d'un système sans être en mesure de parfaitement la caractériser ?

L'alternative que nous utiliserons est simple : on injecte une dynamique. Si nous pouvons prédire le comportement d'un processus dynamique sur un vrai système, nous pouvons conclure avoir bien pris en compte l'organisation de ce système ou, du moins, avoir reproduit les propriétés importantes quant à la dynamique utilisée.

Dans le cas présent, nous appliquerons nos travaux de modélisation à ce qui est couramment appelé l'*épidémiologie sur réseaux*. Le but de cette discipline est d'utiliser la structure en réseau pour mieux décrire la propagation d'une épidémie dans une population humaine. Plus généralement, on peut simplement parler de la propagation d'une information (épidémies, virus informatiques, rumeurs, polluants) dans un ensemble d'éléments (populations humaines, systèmes informatiques, réseaux sociaux ou toiles alimentaires).

Nous utiliserons à cet effet un modèle classique de propagation de maladies semblables à la grippe : la *dynamique SIS*. Cette dynamique implique que les individus sont caractérisés par un état binaire, soit *susceptible* (S), soit *infectieux* (I). Toute la dynamique est centrée sur les individus infectieux qui peuvent infecter leurs voisins susceptibles à un taux $\tau$, mais aussi se rétablir à un taux $r$ et ainsi redevenir susceptible.

Comme nous le verrons au Chap. 4, les modèles d'épidémies sur réseaux utilisent typiquement le paradigme du réseau aléatoire. Ils emploient donc le modèle canonique de réseaux aléatoires présentés dans ce chapitre et par conséquent, ne caractérisent un système que par sa distribution en degré. Pour illustrer l'importance de mieux prendre en compte l'organisation des réseaux complexes, simulons une dynamique SIS sur un réseau humain, le réseau du *cond-mat arXiv* utilisé à la section précédente, et sur son équivalent aléatoire construit à partir du modèle canonique.

À la Fig. 1.14, on présente les résultats de ces simulations. La différence entre les deux courbes illustre clairement l'effet de la topologie (i.e. de l'organisation des liens). Comme le réseau aléatoire est constamment à environ 10% des résultats sur le vrai réseau, il est évident qu'un formalisme analytique utilisant le paradigme du réseau aléatoire ne pourra faire mieux.

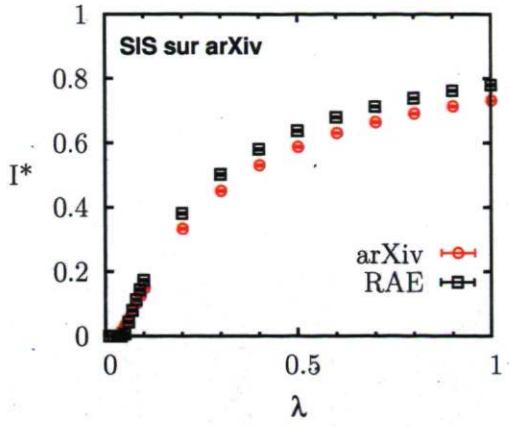



FIG. *1.14:* **Épidémie de grippe sur *arXiv*.** Résultats des simulations d'un modèle SIS de propagation d'épidémie sur le réseau du *cond-mat arXiv* et sur des réseaux aléatoires équivalents (RAE). Les RAE sont produits à partir du modèle canonique de réseau aléatoire présenté à la Sec. 1.2.3. Les résultats correspondent à la fraction infectée du réseau ($I^*$) lorsque la maladie atteint son équilibre entre infections et rétablissements en fonction du ratio du taux d'infection $\tau$ et du taux de rétablissement $r$ ($\lambda = \tau/r$). La différence entre les deux expériences correspond à l'effet de la structure du réseau réel.

Il n'y a ensuite aucune façon de savoir si l'on peut faire confiance à de tels formalismes lorsque vient le temps d'estimer l'effet d'une intervention pour freiner l'épidémie.

Ainsi, nous pourrons non seulement utiliser l'épidémiologie sur réseaux pour valider notre description de l'organisation de réseaux réels, mais également utiliser notre description de réseaux réels pour améliorer les formalismes de prédiction d'épidémies.

### 1.5.2 Vers une description simplifiée

Il est relativement facile de développer un modèle de création de liens résultant en systèmes libres d'échelle, modulaires ou auto-similaires. Cependant, il est beaucoup moins évident de reproduire toutes ces trois propriétés en utilisant le moins de suppositions possibles sur le système. Après tout, plus le modèle est simple, plus les hypothèses sur lesquelles il se base semblent solides. De plus, en s'assurant d'avoir un modèle aussi simple que possible, on conserve la possibilité de décrire analytiquement la structure obtenue. Ceci s'avérera incroyablement utile lorsque vient le temps d'appliquer ce modèle d'organisation à des problèmes dynamiques.

Les prochains chapitres s'attaqueront donc à ces tâches dans cet ordre : développer un modèle d'organisation, le décrire analytiquement et l'utiliser dans un cadre dynamique (modèle SIS). Nous verrons ensuite, en conclusion, si le but visé a été atteint.

# Chapitre 2

# L'attachement préférentiel structurel : les réseaux au-delà du lien

**Structural preferential attachment : Network organization beyond the link**

Laurent Hébert-Dufresne, Antoine Allard, Vincent Marceau,
Pierre-André Noël et Louis J. Dubé
Département de Physique, de Génie Physique, et d'Optique,
Université Laval, Québec, Québec, Canada G1V 0A6.

Référence : Physical Review Letters, 107 (2011), p. 158702.

© 2011 The American Physical Society

## 2.1 Avant-propos

Ce premier chapitre de recherche originale introduit notre nouvelle perspective sur l'organisation des réseaux complexes. Les propriétés universelles, mentionnées au premier chapitre, sont d'abord présentées en les inter-reliant sous un même principe unificateur, avant de développer un nouveau modèle de croissance basé sur cette philosophie. Par la suite, en annexe, on présente des compléments d'informations nécessaires à la reproduction des résultats présentés. Il s'agit en fait d'extraits du matériel supplémentaire offert au lecteur de l'article d'origine.

## 2.2 Résumé

Nous introduisons un mécanisme qui modélise l'émergence des propriétés universelles des réseaux complexes, telles que l'*indépendance d'échelle*, la *modularité* et l'*auto-similarité*, et qui les unifie sous une organisation libre d'échelle au-delà du lien. Pour ce faire, on présente une nouvelle perspective d'organisation des réseaux complexes où les communautés, plutôt que les liens, sont les unités fondamentales des systèmes complexes. Nous montrons comment notre modèle simple peut reproduire des réseaux sociaux et d'information en prédisant leur structure communautaire et plus encore, comment leurs noeuds/communautés sont interconnectés, souvent de façon auto-similaire.

## 2.3 Abstract

We introduce a mechanism which models the emergence of the universal properties of complex networks, such as *scale independence*, *modularity* and *self-similarity*, and unifies them under a scale-free organization beyond the link. This brings a new perspective on network organization where communities, instead of links, are the fundamental building blocks of complex systems. We show how our simple model can reproduce social and information networks by predicting their community structure and more importantly, how their nodes or communities are interconnected, often in a self-similar manner.

## 2.4 A universal matter.

Reducing complex systems to their *simplest possible form* while retaining their important properties helps model their behavior independently of their nature. Results obtained via these abstract models can then be transferred to other systems sharing a similar simplest form. Such

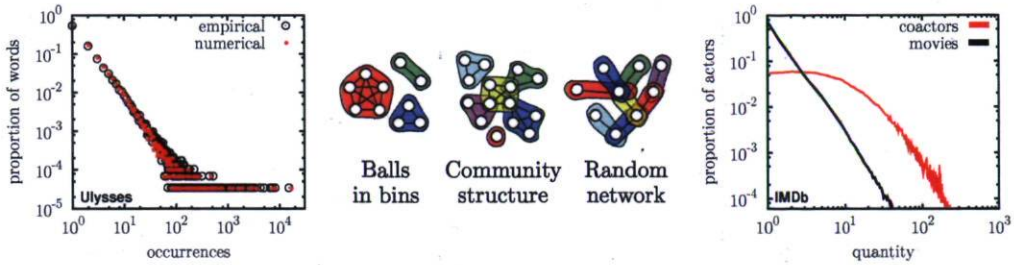
groups of analog systems are called *universality classes* and are the reason why some models apply just as well to the sizes of earthquakes or solar flares than to the sales number of books or music recordings [54]. That is, their statistical distributions can be reproduced by the same mechanism : *preferential attachment*. This mechanism has been of special interest to network science [8] because it models the emergence of power-law distributions for the number of links per node. This particular feature is one of the universal properties of network structure [9], alongside modularity [51] and self-similarity [74]. Previous studies have focused on those properties one at a time [9; 51; 74; 2; 27; 75], yet a unified point of view is still wanting. In this Letter, we present an overarching model of preferential attachment that unifies the universal properties of network organization under a single principle.

Preferential attachment is one of the most ubiquitous mechanisms describing how elements are distributed within complex systems. More precisely, it predicts the emergence of *scale-free* (power-law) distributions where the probability $P_k$ of occurrence of an event of order $k$ decreases as an inverse power of $k$ (i.e., $P_k \propto k^{-\gamma}$ with $\gamma > 0$). It was initially introduced outside the realm of network science by Yule [83] as a mathematical model of evolution explaining the power-law distribution of biological genera by number of species. Independently, Gibrat [22] formulated a similar idea as a law governing the growth rate of incomes. Gibrat's law is the sole assumption behind preferential attachment : the growth rates of entities in a system are proportional to their size. Yet, preferential attachment is perhaps better described using Simon's general *balls-in-bins* process [73].

Simon's model was developed for the distribution of words by their frequency of occurrence in a prose sample [84]. The problem is the following : what is the probability $P_{k+1}(i+1)$ that the $(i+1)$-*th* word of a text is a word that has already appeared $k$ times ? By simply stating that $P_{k+1}(i+1) \propto k \cdot P_k(i)$, Simon obtained the desired distribution [Fig. 2.1(a)]. In this model, the nature of the system is hidden behind a simple logic : the "popularity" of an event is encoded in its number of past occurrences. More clearly, a word used twice is 2 times more likely to reappear next than a word used once. However, before its initial occurrence, a word has appeared exactly zero times, yet it has a certain probability $p$ of appearing for the very first time. Simon's model thus produces systems whose distribution of elements falls as a power law of exponent $\gamma = (2-p)/(1-p)$.

## 2.5 On the matter of networks.

Networks are ensembles of potentially linked elements called *nodes*. In the late 1990s, it was found that the distribution of links per node (the *degree distribution*) featured a power-law tail for networks of diverse nature. To model these so-called *scale-free networks*, Barabási and Albert [9] introduced preferential attachment in network science. In their model, nodes




(a) (b) (c)

FIG. 2.1: **Spectrum of scale-free complex systems.** (a) The distribution of words by their number of appearances in James Joyce's Ulysses (empirical data). The numerical data was obtained from a single realization of Simon's model with $p$ equal to the ratio of unique words (30 030) on the total word count (267 350). (b) Schematization of the systems considered in this Letter, illustrating how order (Simon's model of balls in bins) and randomness (Barabási-Albert's model of random networks) coexist in a spectrum of complex systems. (c) The distribution of coactors and movies per actor in the Internet Movie Database since 2000. The organization moves closer to a true power law when looking at a higher structural level (i.e., movies versus coactors).

are added to the network and linked to a certain number of existing nodes. The probability that the new node chooses an old one of degree $k$ is proportional to $k \cdot N_k$, where $N_k$ is the number of nodes of degree $k$. As the system goes to infinity, $N_k$ falls off as $k^{-3}$.

From the perspective of complex networks, Simon's model may be regarded not as a scheme of throwing balls (e.g., word occurrences) in bins (e.g., unique words), but as an extreme case of scale-free networks where all links are shared within clearly divided structures. Obviously, both Simon's and the Barabási-Albert's (BA) models follow the preferential attachment principle. However, Simon's model creates distinct growing structures, whereas the BA model creates overlapping links of fixed size. By using the same principle, one creates order while the other creates randomness [Fig.2.1(b)]. Our approach explores the systems that lie in between.

## 2.6 When structure matters.

The vast majority of natural networks have a *modular topology* where links are shared within dense subunits [51]. These structures, or *communities*, can be identified as social groups, industrial sectors, protein complexes or even semantic fields [62]. They typically overlap with each other by sharing nodes and their number of neighboring structures is called their *community degree*. This particular topology is often referred to as *community structure* [Fig. 2.1(b)]. Because these structures are so important on a global level, they must influence local

growth. Consequently, they are at the core of our model.

The use of preferential attachment at a higher structural level is motivated by three observations. First, the number of communities an element belongs to, its *membership* number, is often a better indicator of its activity level than its total degree. For instance, we judge an actor taking part in many small dramas more active than one cast in a single epic movie as one of a thousand extras, as we may consider a protein part of many complexes more functional than one found in a single big complex.

Second, studies have hinted that Gibrat's law holds true for communities within social networks [70]. The power-law distribution of community sizes recently observed in many systems (e.g., protein interaction, word association and social networks [62] or metabolite and mobile phone networks [1]) supports this hypothesis.

Third, degree distributions can deviate significantly from true power laws, while higher structural levels might be better suited for preferential attachment models [Fig. 2.1(c)].

## 2.7 A simple model.

Simon's model assigns elements to structures chosen proportionally to their sizes, while the BA model creates links between elements chosen proportionally to their degree. We thus define *structural preferential attachment* (SPA), where both elements and structures are chosen according to preferential attachment. Here, links will not be considered as a property of two given nodes, but as part of structures that can grow on the underlying space of nodes and eventually overlap.

Our model can be described as the following stochastic process. At every time step, a node joins a structure. The node is a new one with probability $q$, or an old one chosen proportionally to its membership number with probability $1 - q$. Moreover, the structure is a new one of size $s$ with probability $p$, or an old one chosen among existing structures proportionally to their size with probability $1 - p$. These two growth parameters are directly linked to two measurable properties : modularity ($p$) and connectedness ($q$) [Fig. 2.2]. Note that, at this point, no assumption is made on how nodes are linked within structures ; our model focuses on the modular organization.

Whenever the structure is a new one, the remaining $s - 1$ elements involved in its creation are once again preferentially chosen among existing nodes. The basic structure size $s$ is called the *system base* and refers to the smallest structural unit of the system. It is not a parameter of the model *per se*, but depends on the considered system. For instance, the BA model directly

creates links, i.e. $s = 2$ (with $p = q = 1$), unlike Simon's model which uses $s = 1$ (with $q = 1$). All the results presented here use a node-based representation ($s = 1$), although they can equally well be reproduced via a link-based representation ($s = 2$). In fact, for sufficiently large systems, the distinction between the two versions seems mainly conceptual (see 3 for details).

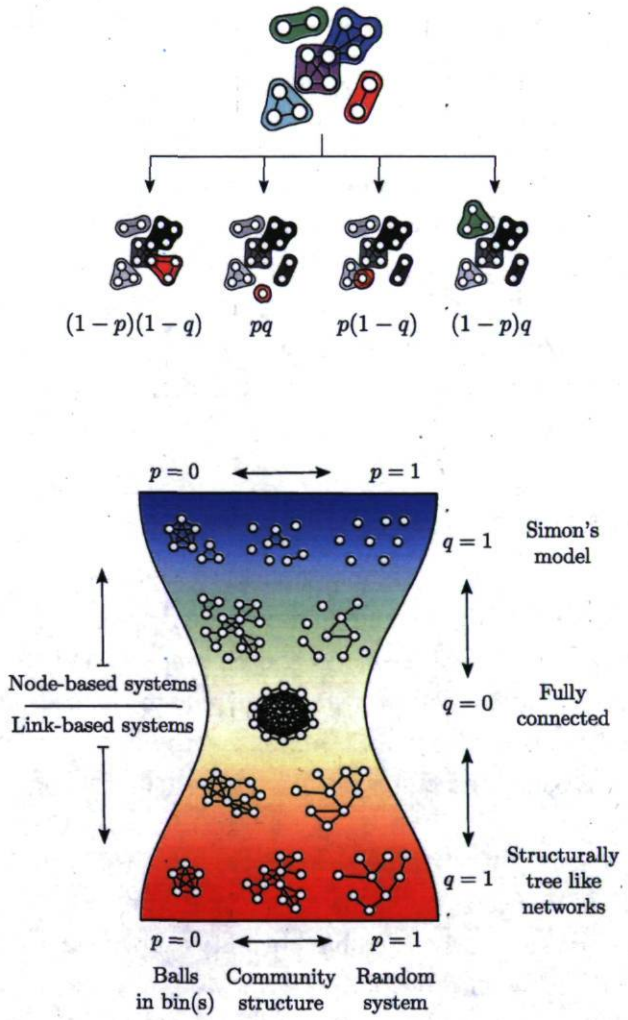



FIG. 2.2: **Structural preferential attachment and the systems it creates.** (top) Representation of the possible events in a step of node-based SPA ; the probability of each event is indicated beneath it. (bottom) A schematization of the spectrum of systems obtainable with SPA. Here, we illustrate the conceptual differences between node-based $s = 1$ and link-based systems $s = 2$ : Simon's model ($q = 1$) creates structures of size one (nodes), while the BA model ($p = q = 1$) creates random networks through structures of size two (links).

In our process, the growth of structures is not necessarily dependent on the growth of the network (i.e., the creation of nodes). Consequently, we can reproduce statistical properties

of real networks without having to consider the large-size limit of the process. This allows our model to naturally include finite size effects (e.g., a distribution cutoff) and increases freedom in the scaling properties. In fact, we can follow $S_n$ and $N_m$, respectively, the number of structures of size $n$ and of nodes with $m$ memberships, by writing master equations for their time evolution [15] :

$$\dot{S}_n(t) = (1-p)\frac{(n-1)S_{n-1}(t) - nS_n(t)}{[1+p(s-1)]\,t} + p\delta_{n,s}\ ; \tag{2.1}$$

$$\dot{N}_m(t) = (1+p(s-1)-q)\frac{(m-1)N_{m-1}(t) - mN_m(t)}{[1+p(s-1)]\,t} + q\delta_{m,1}\ . \tag{2.2}$$

Equations (2.1) and (2.2) can be transformed into ODEs for the evolution of the distribution of nodes per structure and structure per node by normalizing $S_n$ and $N_m$ by the total number of structures and nodes, $pt$ and $qt$, respectively. One then obtains recursively the following solutions for the normalized distributions at statistical equilibrium, $\{\mathcal{S}_n^*\}$ and $\{\mathcal{N}_m^*\}$ :

$$\mathcal{S}_n^* = \frac{\prod_{k=s}^{n-1} k\Omega_s}{\prod_{k=s}^{n}(1+k\Omega_s)} \quad \text{where} \quad \Omega_s = \frac{1-p}{1+p(s-1)} \tag{2.3}$$

$$\mathcal{N}_m^* = \frac{\prod_{k=1}^{m-1} k\Gamma_s}{\prod_{k=1}^{m}(1+k\Gamma_s)} \quad \text{where} \quad \Gamma_s = \frac{1+p(s-1)-q}{1+p(s-1)}\ , \tag{2.4}$$

which scale as indicated in Table 3.1, $\mathcal{N}_m^* \propto m^{-\gamma_N}$ and $\mathcal{S}_n^* \propto n^{-\gamma_S}$.

| System base $s$ | Membership scaling $\gamma_N$ | Size scaling $\gamma_S$ |
|---|---|---|
| Node ($s=1$) | $(2-q)\,/\,(1-q)$ | $(2-p)\,/\,(1-p)$ |
| Link ($s=2$) | $[2\,(p+1)-q]\,/\,(1+q-p)$ | $2/\,(1-p)$ |

TAB. *2.1:* **Scaling exponents of SPA.** Exponents of the power-law distributions of structures per element (membership) and of elements per structure (size) at statistical equilibrium. One easily verifies that the membership scaling of link-based systems with $p=q=1$ corresponds to that of the BA model ($\gamma_N = 3$), and that node-based systems with $q = 1$ reproduce Simon's model. See Chap. 3 for the derivation.

## 2.8 Results and discussions.

There are three distributions of interest which can be directly obtained from SPA : the membership, the community size, and the community degree distributions. In systems such as the size of business firms or word frequencies, these distributions suffice to characterize the organization. To obtain them, the SPA parameters, $q$ and $p$, are fitted to the empirical scaling exponents of the membership and community size distributions. In complex networks, one may also be interested in the degree distribution. Additional assumptions are then needed to determine how nodes are interconnected within communities (specified when required).

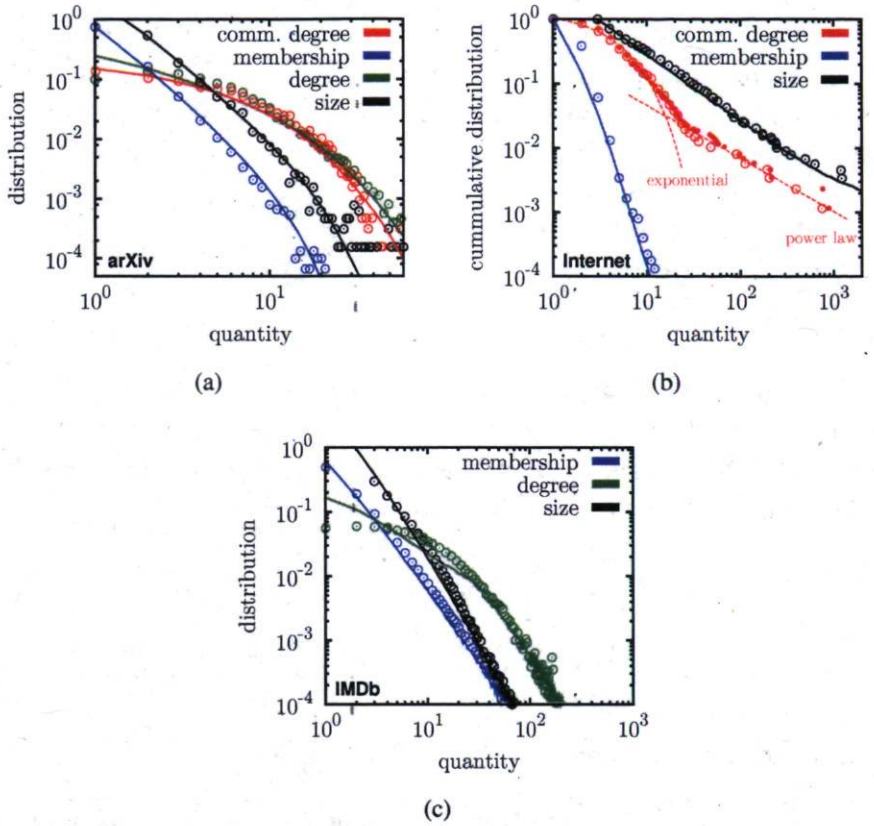



FIG. 2.3: **Reproduction of real systems with SPA.** Circles : distributions of topological quantities for (a) the *cond-mat arXiv* circa 2005 ; (b) Internet at the level of autonomous systems circa 2007 ; (c) the IMDb network for movies released since 2000. Solid lines : average over multiple realizations of the SPA process with (a) $p = 0.56$ and $q = 0.59$ ; (b) $p = 0.04$ and $q = 0.66$ ; (c) $p = 0.47$ and $q = 0.25$. For each realization, iterations are pursued until an equivalent system size is obtained. The Internet data highlights the transition between exponential and scale-free regimes in a typical community degree distribution. It is represented by a single realization of SPA (dots), because averaging masks the transition.

The first set of results considered is the community structure of the coautorship network of an electronic preprints archive, the *cond-mat arXiv* circa 2005 [Fig. 2.3(a)], whose topology was already characterized using a clique percolation method [62]. Here, the communities are detected using the link community algorithm of Ahn *et al.* [1], confirming previous results.

Using only two parameters, our model can create a system of similar size with an equivalent topology according to the four distributions considered (community sizes, memberships, community degree and node degree). Not only does SPA reproduce the correct density of structures of size 2, 3, 4 or more, but it also correctly predicts *how* these structures are interconnected via their overlap, i.e., the community degree. This is achieved without imposing

any constraints whatsoever for this property. The first portion of the community degree distribution is approximately exponential ; a behavior which can be observed in other systems, such as the Internet [Fig. 2.3(b)] and both a protein interaction and a word-association network [62]. To our knowledge, SPA is the first growth process to reproduce such community structured systems.

Moreover, assuming fully connected structures, SPA correctly produces a similar behavior in the degree distribution of the nodes. Obtaining this distribution alone previously required two parameters and additional assumptions [2]. In contrast, SPA shows that this is a signature of a *scale-free community structure*. This is an interesting result in itself, since most observed degree distributions follow a power law only asymptotically. Furthermore, this particular result also illustrates how self-similarity between different structural levels (i.e., node degree and community degree distributions) can emerge from the scale-free organization of communities.

Finally, the Internet Movie Database coacting network is used to illustrate how, for bigger and sparser communities which cannot be considered fully connected, one can still easily approximate the degree distribution. We first observe that the mean density of links in communities of size $n$ approximately behaves as $\log(n)/n$ (see Supplemental Material, Sec. 2.10). Then, using a simple binomial approximation to connect the nodes within communities, it is possible to approximate the correct scaling behavior for the degree distribution [Fig. 2.3(c)]. This method takes advantage of the fact that communities are, by definition, homogeneous such that their internal organization can be considered random.

## 2.9 Conclusion and perspective.

In this Letter, we have developed a complex network organization model where connections are built through growing communities, whereas past efforts typically tried to arrange random links in a scale-free, modular and/or self-similar manner. Our model shows that these universal properties are a consequence of preferential attachment at the level of communities : the scale-free organization is inherited by the lower structural levels.

Looking at network organization beyond the link is also useful to account for missing links [13] or to help realistic modeling [30; 36]. For instance, this new paradigm of scale-free community structure suggests that nodes with the most memberships, i.e., structural hubs, are key elements in propagating epidemics on social networks or viruses on the Internet. These structural hubs connect many different neighborhoods, unlike standard hubs whose links can be redundant if shared within a single community.

There is no denying that communities can interact in more complex ways through time [61]. Yet, from a statistical point of view, those processes can be neglected in the context of a structurally preferential growth. Similarly, even though other theories generating scale-free designs exist [16], they could also benefit from generalizing their point of view to higher levels of organization.

## 2.10 Appendix : Excerpts of Supplemental Material

This section gives more details on the datasets used in the Letter and on the methods employed to characterize their topology.

### 2.10.1 Data

**Internet Movie Database** The dataset used for the co-acting network of IMDb consists only of movies released after December 31rst 1999. Interestingly, the degree distribution is almost identical to that published a decade earlier [9] which consisted of all movies released before the turn of the century. This suggests, since the two networks contain distinct and exclusive ensembles of movies, that the growth parameters of the IMDb network are constant. The network contains 7 665 259 links between 716 463 nodes (actors), where two actors share a link if they are credited alongside another for at least one movie. It was only analysed using the link community algorithm, because of memory issues with CFinder. The organization levels corresponding to actual movies, which is how the dataset was originally compiled, was deemed unsuitable for the study because of the presence of economic (limiting the number of actors in a movie) and artistic (typically requiring a minimal number of characters in a movie) constraints. We believe that a community detection process on the network actually frees the system from these constraints and yield communities of actors linked by genre, time, location, etc.

**arXiv** 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 ; resulting in a network of 125 959 links between 30 561 nodes (authors). This dataset was compiled, analysed and presented in [62].

**Internet** 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 and was not

previously published. The network contains 22 962 nodes and 48 436 links.

### 2.10.2 Levels of organization

The first step when looking to compare the structure of real networks with systems produced by SPA is to analyse the empirical data. As mentionned earlier, our main algorithm (the link community algorithm [1]) has a single parameter to tune for community detection : its Jaccard threshold. The Jaccard threshold embodies *how similar* the neighborhoods of the ends of two links must be in order for these links to be considered as part of the same link community. Tuning this parameter, demanding how tightly connected a group of nodes must be in order to be labeled as a community, allows us to look at different levels of organization within the network. If too small, the algorithm will most likely end up with communities corresponding to the connected components of the networks. If too big, significant communities will be broken up into different smaller ones. In this paper, we proceeded by sweeping this parameter in order to find the level of scale-free organization.

### 2.10.3 A note on node-based and link-based systems

All results presented in this work used a node-based version of SPA. Which means that new structures contain a single node and that they will remain disconnected from the other components of the network until they reach an older node. For the IMDb data, this choice is not even a question as the network contains many such satellites structures (even some of size one) which are disconnected from the giant component. In other systems, like the arXiv network, the choice can be more complicated. One might be tempted to use a link-based system process to reproduce the arXiv, since it is a co-author network and thus cannot contain isolated nodes. However, it does contain some disconnected components, which a link-based process like the Barabási-Albert model [9] is incapable of producing. Hence, it seemed logical to use the node-based process and simply remove the structures of size one (nodes who failed to co-author a paper) from the final system.

As a final point on the subject, it is interesting to note that we have been able to reproduce all results using both the node-based and link-based version of SPA. In sufficiently large and connected systems, the distinction between the two seems mainly conceptual.

### 2.10.4 Supplementary results and discussions

The Letter presented our results for the arXiv network, the Internet and the Internet Movie Database. The arXiv data is completely shown, but the Internet is illustrated for communities

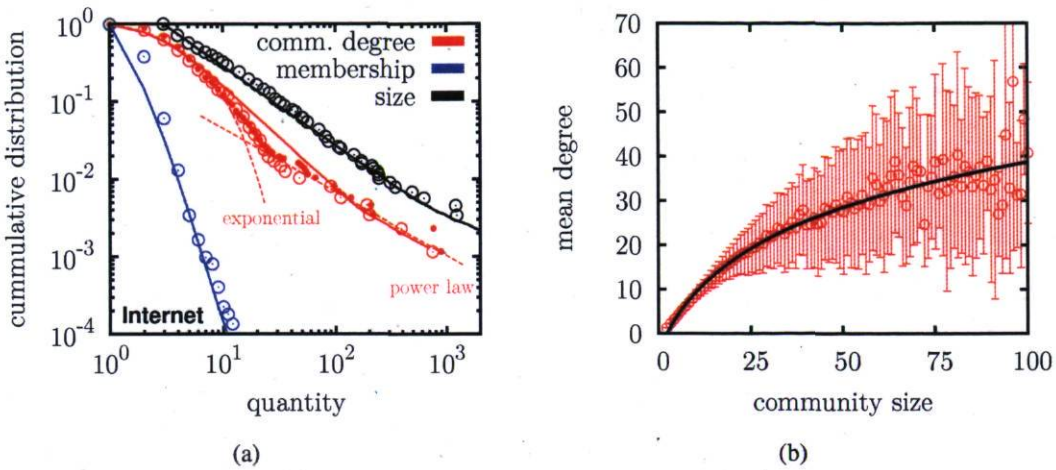



FIG. 2.4: **Community structure and structural preferential attachment (SPA).** (a) ⊙ : distributions of topological quantities for the ensemble of the Internet at the level of autonomous system circa 2007 ; solid lines : average over multiple realizations of the SPA process with $p = 0.04$ and $q = 0.66$. The empirical network was analysed using the link community algorithm [1] with Jaccard threshold 0.08. (b) The mean number of links per node within a given community as a function of the community size in the IMDb network. The fit is done using a logarithmic function of the form $f(x) = a \cdot log(x+b) - c$.

of size 3 or bigger (as done by the authors of the detection algorithm [1]) because the algorithm can overestimate the number of communities of size 2 and the goal is here to highlight the connectedness of communities. For the IMDb, the community size distribution is normalized for communities of size 3 or bigger, but the communities of size 2 are considered in the membership and degree distributions. These results highlight how these systems follow a scale-free community structure and how SPA can be used to predict behavior *outside* of the model's specification. More precisely, the numerical systems predict how the communities are interconnected via their overlap, reproducing the exponential behavior and the heavy tail of the community degree distribution.

It is interesting to note that averaging over many iterations of the SPA process highlights the distribution cut-off caused by the finite size of the system. This effect is mostly visible on the arXiv results. On the other hand, because the position of the transition between exponential and power-law behavior observed in the cumulative community degree distribution is highly dependent on the amount of "leading" structures (i.e. the number of structures which are able to break away from the majority and thus have a significantly bigger size), it can differ slightly between two realizations of SPA. In this context, averaging over multiple iterations of the same process partly smooths out the transition. For this reason, a single realization of the model is also presented on Fig. 2.4(a) to better illustrate the behavior of community degree in a finite system.

### 2.10.5 From communities, back to links

This last subsection presents results which, although preliminary, imply that individuals within a given social community can be approximated as being randomly connected.

The first step in shifting our point of view from communities back to links is to evaluate just how connected the communities of our systems are. Figure 2.4(b) illustrates the mean number of links per node within a given community as a function of the community size, which is found to grow logarithmically. Using this measure to determine the density of a structure of a given size, we simply throw a dice for each possible link to determine which links actually exist, while respecting the actual density of the network. This allows us to move from a potential degree distribution to an estimated degree distribution. If the binomial approximation (all links within a given community exist only with a certain probability) is correct, this estimated degree distribution should be close to the actual degree distribution of the system we are trying to reproduce. According to Fig. 2.3(c), this is indeed the case. It is easy to note that the number of nodes of small degree is overestimated by SPA. This is either a consequence of SPA producing too many small satellite components around the main network, or a consequence of IMDb sampling method, where an actor who has only acted in a small scale short film with one or two co-actors is more likely to be absent from the database than actors with hundreds of co-actors.

# Chapitre 3

# Croissance de systèmes libres d'échelle, modulaires et auto-similaires

**Structural preferential attachment :**
**Stochastic process for the growth of scale-free, modular and self-similar systems**

Laurent Hébert-Dufresne, Antoine Allard, Vincent Marceau,
Pierre-André Noël et Louis J. Dubé
Département de Physique, de Génie Physique, et d'Optique,
Université Laval, Québec, Québec, Canada G1V 0A6.

arXiv :1109.0034

## 3.1 Avant-propos

Ce chapitre est en fait la version prépublication d'un tout nouvel article. Alors que cet article peut être vu comme un complément du précédent, il contient deux contributions principales s'insérant dans un contexte plus large que l'attachement préférentiel structurel. Premièrement, on y présente des méthodes d'analyse s'appliquant aussi bien à l'attachement préférentiel classique qu'à notre nouveau modèle d'attachement préférentiel structurel. Deuxièmement, le résultat principal de l'article concerne une propriété statistique des systèmes libres d'échelle ne s'appliquant pas, a priori, aux réseaux complexes[1]. Quoi qu'il en soit, la lecture de cet article clarifie les idées présentées au chapitre précédent, en plus d'apporter de nouveaux résultats. Les lecteurs intéressés devraient donc y trouver matière à réflexion.

## 3.2 Résumé

Il est connu que plusieurs systèmes complexes partagent des propriétés universelles d'organisation, telles que l'*indépendance d'échelle*, la *modularité* et l'*auto-similarité*. Nous empruntons des outils de la physique statistique dans le but d'étudier l'*attachement préférentiel structurel* (APS), un principe de croissance récemment proposé pour modéliser l'émergence des propriétés mentionnées. Cet article étudie le processus stochastique correspondant en termes de son évolution temporelle, de son comportement asymptotique et des propriétés d'échelle de son équilibre statistique. De plus, certaines approximations sont introduites dans le but de faciliter la reproduction de systèmes réels, principalement de réseaux complexes, avec l'APS. Finalement, nous nous intéressons à un comportement particulier observé dans le processus stochastique, la *dynamique de peloton*, et montrons qu'il prédit plusieurs propriétés de vrais systèmes en croissance. On confirme les prédictions en prenant comme exemples des extraits de prose.

## 3.3 Abstract

Many complex systems have been shown to share universal properties of organization, such as *scale independence*, *modularity* and *self-similarity*. We borrow tools from statistical physics in order to study *structural preferential attachment* (SPA), a recently proposed growth principle for the emergence of the aforementioned properties. We study the corresponding stochastic process in terms of its time evolution, its asymptotic behavior and the scaling

[1]Nous verrons que cette propriété n'émerge que sous le moyennage de plusieurs réalisations du même système. Or, la grande majorité des réseaux complexes ont le malheur d'être unique !

properties of its statistical steady state. Moreover, approximations are introduced to facilitate the modelling of real systems, mainly complex networks, using SPA. Finally, we investigate a particular behavior observed in the stochastic process, the *peloton dynamics*, and show how it predicts some features of real growing systems using prose samples as an example.

## 3.4 Introduction

In a recent contribution, we have proposed a model of network organization [29] based on a generalization of the classical preferential attachment principle (PA) [72; 9] to a higher order : structural preferential attachment (SPA). In this model, elements of the system join and create structures. In all attachment events, both the element and the structure involved are chosen proportionally to their past activities. Elements can represent money being invested, written words, individuals in a social network, proteins or websites, while the structures can be business firms, semantic fields, friendships and communities, protein complexes or types of activities and interest [72; 9; 62; 1].

SPA can be described by the following stochastic process (see Fig. 3.1 for a visual aid). At every time step, an element joins a structure. With probability $q$, the element is a new one ; or with probability $1 - q$, it is chosen among existing elements proportionally to the current number of structures to which they belong (i.e. their *membership* number). Moreover, with probability $p$, the structure is a new one of size $s$ ; or with probability $1 - p$, it is chosen among existing structures proportionally to the current number of elements they possess (i.e. their *size*). Whenever the structure is a new one, the remaining $s - 1$ elements involved in its creation are once again preferentially chosen among existing nodes. The basic structure size $s$ is called the *system base* and refers to the smallest structural unit of the system. For example, if $s = 1$, the system base is simply the elements themselves and we refer to this version as *node-based* SPA, while if $s = 2$, the system base is a pair of elements resulting in *link-based* SPA.

This stochastic process can either be seen as a scheme of throwing balls (the elements) in bins (the structures) or as a process of network growth. In the latter, the elements are the *nodes* of the network while the structures represent significant topological patterns, motifs, modules or *communities*, within which elements are linked.

SPA results in the growth of *modular* systems, because modules (or structures) are the basic building blocks of the model. These systems are also *scale-free*, in the sense that their main statistical features (membership and size distributions) converge toward power laws (free of any characteristic scale) as a result of the preferential attachment principle [72; 9]. Finally, these systems are said to be *self-similar* as different levels of organization follow the

same general behavior : elements are interconnected with one another by sharing structures in the same way the structures themselves are interconnected by sharing elements.

In this paper, we borrow tools from statistical physics to study SPA in detail. In Sec. 3.5, an exact description of SPA is obtained by writing the corresponding discrete stochastic process. From this description, we obtain the statistical steady-state of the resulting system with asymptotic expressions of its scaling behaviors. In Sec. 3.6, some useful approximations are introduced and studied in order to facilitate the comparison between systems produced by SPA and real-world systems, using the *cond-mat arXiv* co-author network as an example. In order to investigate the validity of these approximations, we then study the existence of correlations between elements and structures, in both the SPA process and in the *cond-mat arXiv*. Lastly, in Sec. 3.7, we highlight an interesting behavior of discrete PA processes, which we call the *peloton dynamics*, by comparing the initial stochastic process with an explicit solution for the time evolution of the continuous time version (whose derivation is presented in Appendix). We then seek empirical evidences of this behavior in growing prose samples.

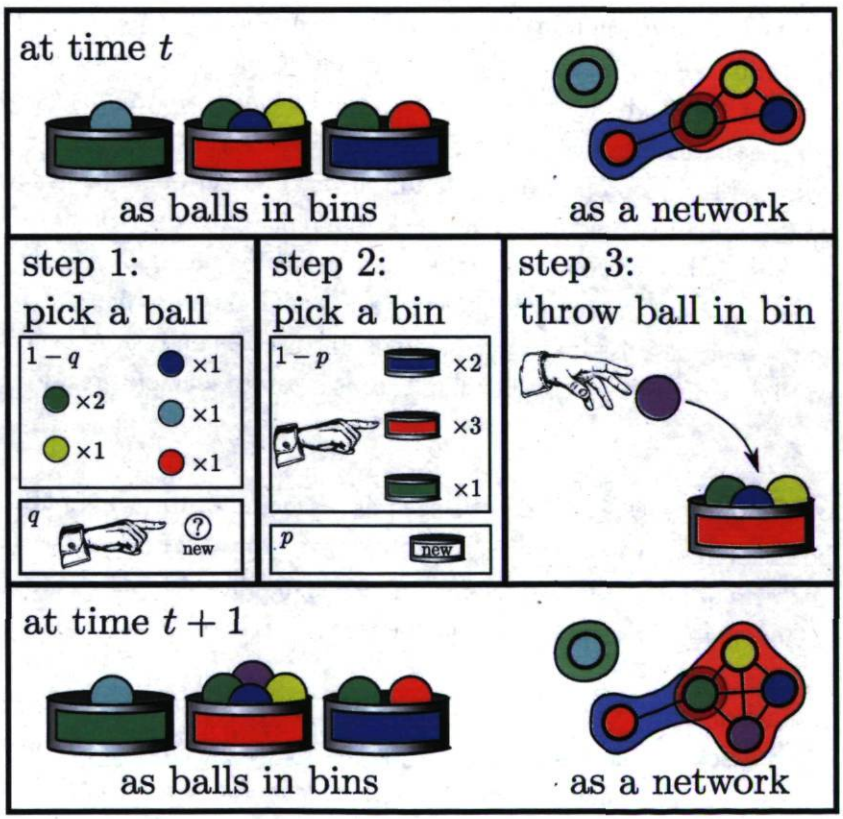



FIG. *3.1:* **A step of node-based SPA.**

## 3.5 Stochastic process

### 3.5.1 Time evolution

To follow the growth of a system as prescribed by the SPA process, we separate elements and structures. We will distinguish nodes by their respective number of memberships, $m$, and structures by their respective size, $n$, as these are the only features relevant to their evolution. Let $\tilde{N}_m(t)$ be the mean number of elements (or *nodes* to use the network terminology) with $m$ memberships and $\tilde{S}_n(t)$ be the mean number of structures of size $n$.

At each time step, the evolution of these quantities is twofold : first, a constant increment for potential new nodes and structures ; second, an operation corresponding to the preferential growth of existing nodes and structures. More clearly, each time step corresponds to an iteration of the following rule :

$$\tilde{N}_m(t+1) = \tilde{N}_m(t) + q\delta_{m,1} + \frac{1-q+p(s-1)}{t[1+p(s-1)]}\left[(m-1)N_{m-1}(t) - mN_m(t)\right] \tag{3.1}$$

$$\tilde{S}_n(t+1) = \tilde{S}_n(t) + p\delta_{n,s} + \frac{1-p}{t[1+p(s-1)]}\left[(n-1)S_{n-1}(t) - nS_n(t)\right] . \tag{3.2}$$

The two increments $q\delta_{m,1}$ and $p\delta_{n,s}$, where $\delta_{i,j}$ is the Kronecker delta, correspond to birth events for elements (with one membership) and structures (of size $s$), respectively. The last increments correspond to the growth of old entities, where an element has a negative effect on itself and a positive effect on its neighbor (e.g. $\tilde{N}_m \rightarrow \tilde{N}_{m+1}$) at a given rate and the denominator $t[1+p(s-1)]$ normalizes the preferential attachment probabilities.

This iterative description is straightforward, yet we can define the system in closed form by using generating functions (GFs) [81]. We define two functions whose power series coefficients correspond to the elements of our two distributions :

$$\tilde{N}(x;t) = \sum_m \tilde{N}_m(t)x^m \quad \text{and} \quad \tilde{S}(x;t) = \sum_n \tilde{S}_n(t)x^n \tag{3.3}$$

where the tildes once again refer to the fact that these functions generate mean numbers of elements and structures. In terms of these GFs, Eqs. (3.1) and (3.2) can be rewritten as :

$$\tilde{N}(x;t+1) = \left(1 + \frac{\Gamma_s}{t}x(x-1)\frac{d}{dx}\right)\tilde{N}(x;t) + qx\,; \tag{3.4}$$

$$\tilde{S}(x;t+1) = \left(1 + \frac{\Omega_s}{t}x(x-1)\frac{d}{dx}\right)\tilde{S}(x;t) + px^s\,, \tag{3.5}$$

where we also introduced

$$\Gamma_s = \frac{1-q+p(s-1)}{1+p(s-1)} \quad \text{and} \quad \Omega_s = \frac{1-p}{1+p(s-1)}\,. \tag{3.6}$$

A similar description can be obtained in terms of the corresponding probability generating functions (PGFs), $\mathcal{N}(x;t)$ and $\mathcal{S}(x;t)$, which generate the distributions of memberships per element and size per structures respectively. To transform the previous description in terms of these PGFs, note that the mean numbers of elements, $\tilde{N}_m$, or structures, $\tilde{S}_n$, in a given state corresponds to the proportion of such elements, $N_m$, or structures, $S_n$, multiplied by the total number of elements, $qt$, or structures, $pt$, at time $t$. One can now rewrite Eqs. (3.4) and (3.5) in terms of $\mathcal{N}(x;t)$ and $\mathcal{S}(x;t)$ by multiplying these functions by $qt$ and $pt$, respectively :

$$(t+1)\,\mathcal{N}(x;t+1) = \left(t+\Gamma_s x(x-1)\frac{d}{dx}\right)\mathcal{N}(x;t)+x \tag{3.7}$$

$$(t+1)\,\mathcal{S}(x;t+1) = \left(t+\Omega_s x(x-1)\frac{d}{dx}\right)\mathcal{S}(x;t)+x^s. \tag{3.8}$$

As we will see in what follows, the description in terms of PGFs is generally more useful and will hereafter be used in our results to validate the analytical description.

### 3.5.2 Degree distributions

PGFs provide simple ways to evaluate secondary properties of a given state. For example, the node degree distribution and the community degree distribution. The former describes how many elements can be reached from a randomly chosen element, in other words the number of links connected to this node in the network representation. The latter refers to a similar concept, namely the number of structures that overlap (by sharing elements) with one randomly chosen structure.

To illustrate how this calculation is performed, one can simply refer to the composition property of PGFs. We first pick a random element, its membership distribution is generated by $\mathcal{N}(x;t)$. For every possible value of its membership number $m$, we must sum over all possible cases for the different sizes of these structures. However, we know that all of these $m$ structures have *at least* one element. It is thus $k$ times more likely that one of these $m$ structures is a structure of size $k$ than a structure of size one. Furthermore, we do not want to count the initial element we chose, and will thus reduce the size of each structure by one. Hence, their size distribution is not generated by $\mathcal{S}(x;t)$, but instead by $\mathcal{S}'(x;t)/\mathcal{S}'(1;t)$, where the denominator acts as a normalisation factor. Knowing that the convolution of two sequences is generated by the product of the corresponding PGFs, one can take the $m$-th power of the new size PGF to obtain the PGF for the sum of $m$ structures. Finally, we sum over all possible values of $m$ to obtain :

$$D(x;t) = \sum_m N_m \left[\mathcal{S}'(x;t)/\mathcal{S}'(1;t)\right]^m = \mathcal{N}\left(\left[\mathcal{S}'(x;t)\Big/\mathcal{S}'(1;t)\right],t\right). \tag{3.9}$$

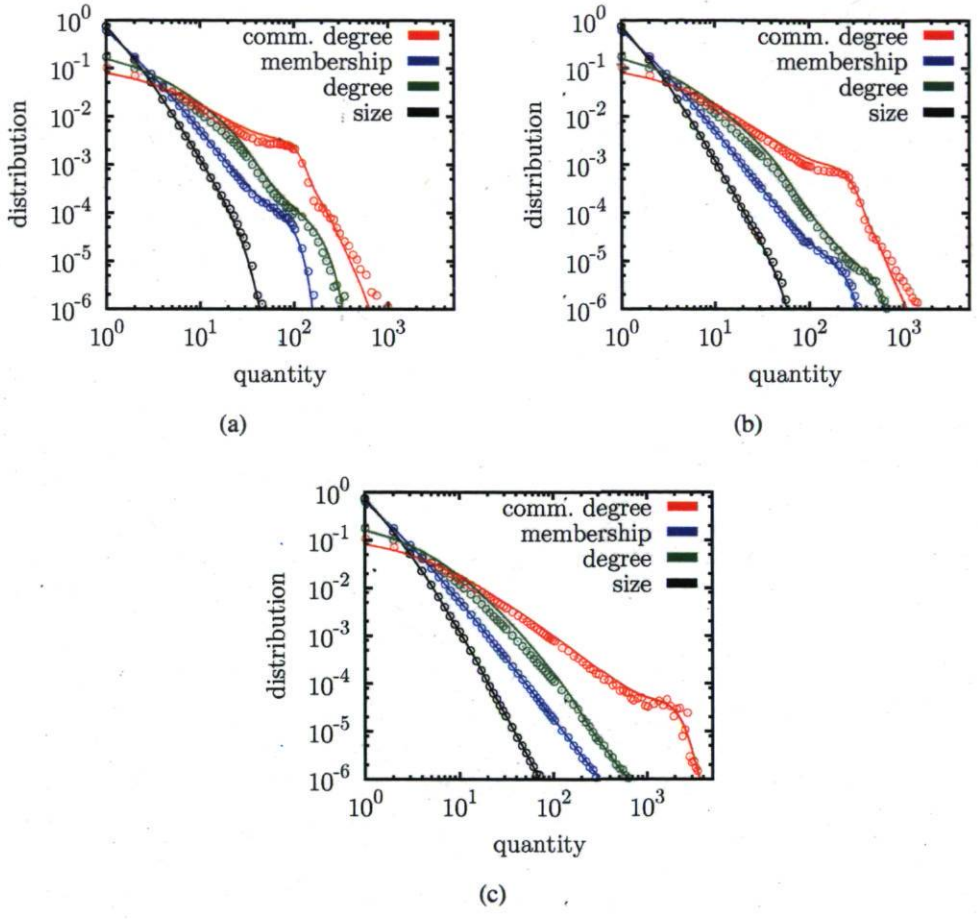



FIG. *3.2:* **Validation of the analytical description of SPA.** Time evolution of node-based SPA using $q = 0.35$ and $p = 0.65$ for the four main characteristics of the topology : memberships, community size, node degree and community degree. Snapshots are taken when the systems reach a) 250 structures, b) 1000 structures and c) 25 000 structures. Shown by markers are Monte-Carlo results averaged over 25 000 simulations ; analytical predictions of Eqs. (3.7 - 3.10) are shown with continuous lines.

Using the same logic for structures and their community degree, one can write :

$$C(x;t) = \mathcal{S}\left(\left[\mathcal{N}'(x;t)\Big/\mathcal{N}'(1;t)\right],t\right) . \tag{3.10}$$

The self-similarity between different levels of organization in the systems created by SPA stems from the similarity between Eqs. (3.9) and (3.10). As long as $\mathcal{N}(x;t)$ and $\mathcal{S}(x;t)$ are similar, the various possible compositions, which represent different organization properties, will also be similar.

The validation of our analytical description for the time evolution of SPA is presented on Fig. 3.2 using Monte-Carlo simulations. Note that our calculations for the degree distributions are merely approximations because they suppose *homogeneous mixing* between elements and structures, while an element with $m = i$ might not see exactly the same size distribution as an element with $m = j$. Such element-structure correlations are investigated in a latter section of the paper.

### 3.5.3 Statistical equilibrium

The statistical equilibrium can be imposed by setting $\mathcal{N}(x;t+1) = \mathcal{N}(x,t) \equiv \mathcal{N}^*(x)$ and $\mathcal{S}(x,t+1) = \mathcal{S}(x,t) \equiv \mathcal{S}^*(x)$ in Eqs. (3.7) and (3.8), yielding :

$$\mathcal{N}^*(x) = \Gamma_s x(x-1)\frac{d}{dx}\mathcal{N}^*(x) + x ; \tag{3.11}$$

$$\mathcal{S}^*(x) = \Omega_s x(x-1)\frac{d}{dx}\mathcal{S}^*(x) + x^s . \tag{3.12}$$

These ordinary differential equations can be solved straightforwardly to obtain their solutions in terms of hypergeometric functions of the form ${}_2F_1\,(a,b;c;x)$ :

$$\mathcal{N}^*(x) = \frac{x}{1+\Gamma_s}\,{}_2F_1\left(1,1;2+\frac{1}{\Gamma_s};x\right) , \tag{3.13}$$

and :

$$\mathcal{S}^*(x) = \frac{(s-1)!\Omega_s^{s-1}x^s}{1+s\Omega_s}\,{}_2F_1\left(1,s;(s+1)+\frac{1}{\Omega_s};x\right) . \tag{3.14}$$

The statistical equilibrium for the two distributions of interest can now be obtained through the power series coefficients of these two functions :

$$N_m^*(s) = \frac{\prod_{k=1}^{m-1} k\Gamma_s}{\prod_{k=1}^{m}(1+k\Gamma_s)} , \quad S_n^*(s) = \frac{\prod_{k=s}^{n-1} k\Omega_s}{\prod_{k=s}^{n}(1+k\Omega_s)} . \tag{3.15}$$

These solutions for the asymptotic behavior of the statistical distributions can be validated through comparison with the long term behavior of our predicted time evolution, as done in Fig. 3.3.

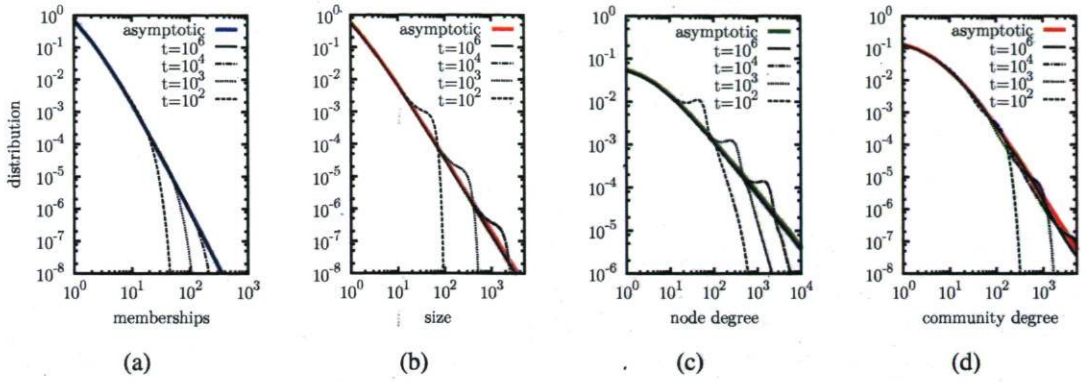



FIG. *3.3:* **Validation of the asymptotic solution of SPA.** Convergence of the time evolution toward equilibrium for a) the membership distribution, b) the size distribution, c) node degree distribution and d) community degree distribution in node-based SPA using $q = 0.6$ and $p = 0.25$.

### 3.5.4 Scaling behaviour

From PA, it is well known that the $N_m^*$ and $S_n^*$ distributions will fall as power laws, i.e.

$$N_m^* \propto m^{-\gamma_N} \quad \text{and} \quad S_n^* \propto n^{-\gamma_S} \,. \tag{3.16}$$

To calculate the $\gamma_N$ scaling exponent, we can evaluate the following ratio using Eq. (3.15)

$$\lim_{m\to\infty} \frac{N_m^*}{N_{m-1}^*} = \lim_{m\to\infty} \left(\frac{m}{m-1}\right)^{-\gamma_N} = \lim_{m\to\infty} \frac{(m-1)\,\Gamma_s}{1+m\Gamma_s} \tag{3.17}$$

from which it follows that

$$\gamma_N = \lim_{m\to\infty} \frac{\log\left((m-1)\,\Gamma_s \Big/ (1+m\Gamma_s)\right)}{\log\left((m-1)\Big/m\right)} = 1 + \frac{1}{\Gamma_s} \,. \tag{3.18}$$

Similarly, one can directly write for structures :

$$\gamma_S = 1 + \frac{1}{\Omega_s} \,. \tag{3.19}$$

The node and community degree distributions, as compositions of two power-law distributions, will fall as the slower of the two original distributions. Noting that $N'(x,t)$ and $S'(x,t)$ will follow $\gamma_{N'} = \gamma_N - 1$ and $\gamma_{S'} = \gamma_S - 1$ because of the derivative, we obtain :

$$\gamma_D = \min\left\{\gamma_N, \gamma_S - 1\right\} \quad \text{and} \quad \gamma_C = \min\left\{\gamma_N - 1, \gamma_S\right\} \,. \tag{3.20}$$

These results are validated on Fig. 3.4.

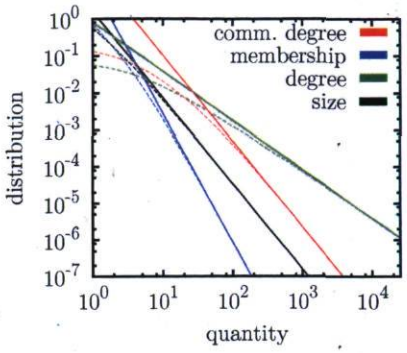



FIG. *3.4:* **Validation of the scaling exponents of SPA.** Validation of Eqs. (3.18) and (3.19) as predictions for the asymptotic scaling behaviours of the main statistical distributions (dotted lines : steady-state solutions, continuous line : scaling predictions) for node-based SPA using $q = 0.6$ and $p = 0.25$.

## 3.6 Approximations and limitations

To complete our description of the SPA process, this section examines some approximations that have either proven useful when reproducing empirical data with the SPA process or that correspond to limitations of the presented formalism.

### 3.6.1 Correspondence between system bases

In [29], we mentionned that the system base $s$ was not a parameter of the model per se, but depended on the information available or on the nature of the system. For instance, the World-Wide Web is mapped by following links between webpages, such that it is impossible to find a page with no links. The smallest structural unit is thus the link and not the webpage itself : it is a link-based system ($s = 2$). Some systems reproduced in [29] with node-based SPA ($s = 1$) were actually link-based, for example the author collaboration network of the *cond-mat arXiv*, where authors only appear once they have at least one collaboration. This was done by ignoring structures of size one when compiling the final system created by the node-based SPA. Furthermore, structures of size one can rarely be detected in network data if they are not completely disconnected from the rest of the systems. Hence, it is useful to be able to ignore these structures at the end of the stochastic growth process, independently of the system base.

For the size distribution, ignoring structures of size one simply implies a renormalization for structures of size two or greater. Noting the PGF for an approximate link-based SPA

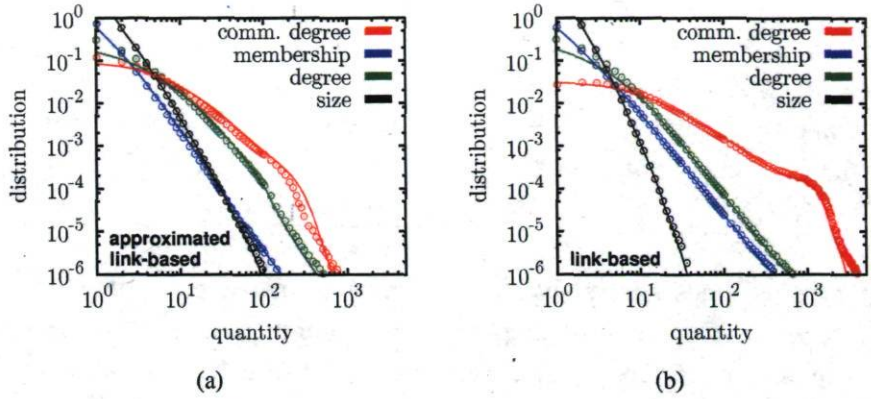



FIG. *3.5:* **Comparison between node-based and link-based SPA.** Analytical predictions and simulations for a) an approximation of link-based system from a node-based SPA process and b) the link-based SPA using the same parameters as the previous figure. Note how the change in system base, with constant parameters, greatly modifies the produced system. This highlights both the validity of Eqs. (3.21 - 3.22) (which feature two levels of approximation of homogeneous mixing) and the importance of considering the influence of the system base on the scaling behaviour (see Eq. (3.24)).

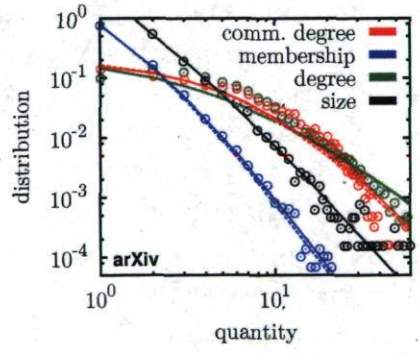



FIG. *3.6:* **Reproducing the arXiv with analytical results.** Community structure of the *cond-mat arXiv* as measured by a link community algorithm [1] (dots) and as modelled by link-based SPA with $q = 0.95$ ; $p = 0.39$ in continuous lines or node-based SPA used to approximate a link-based system with $q = 0.68$ ; $p = 0.56$ according to Eq. (3.24) in dotted lines. The two black lines perfectly overlap, while one membership distribution is shifted by approximation (3.22).

$\mathcal{S}_2^{\text{app}}(x)$ using the original node-based functions $\mathcal{S}_1(x)$, we can write :

$$\mathcal{S}_2^{\text{app}}(x) = \frac{\mathcal{S}_1(x) - S_1 x}{\mathcal{S}_1(1) - S_1} . \tag{3.21}$$

For the membership distribution, once again assuming homogeneous mixing, it means we must randomly remove the fraction of memberships which corresponds to the structures of size one. Using the composition of PGFs, this can be done by composing the membership PGF with the PGF for a binomial trial :

$$\mathcal{N}_2^{\text{app}}(x) = \frac{\mathcal{N}_1(x(1-\epsilon)+\epsilon) - \mathcal{N}_1(\epsilon)}{1 - \mathcal{N}_1(\epsilon)} \tag{3.22}$$

where $\mathcal{N}_1(\epsilon)$ corresponds to the elements left with no memberships and thus need to be removed from the system. This trial will remove a fraction $\epsilon$ of memberships, where $\epsilon$ corresponds to the fraction of memberships which are associated with structures of size one :

$$\epsilon = \frac{S_1}{\sum_n n S_n} = \frac{\mathcal{S}_1'(0)}{\mathcal{S}_1'(1)} . \tag{3.23}$$

The validity of this approximate description and the effects of switching between system bases are illustrated on Fig. 3.5.

To compare the results of approximated and actual link-based SPA for the same community structure, we first need to identify the relation between the parameter pairs $\{q_1, p_1\}$ and $\{q_2, p_2\}$ which is such that $\Gamma_1 = \Gamma_2$ and $\Omega_1 = \Omega_2$. From Eq. (3.6), we obtain :

$$p_2 = \frac{p_1}{2 - p_1} \quad \text{and} \quad q_2 = (1 + p_2) q_1 . \tag{3.24}$$

While it is easily verified that ignoring structures of size one in node-based SPA can result in statistical features similar to that of link-based SPA (see Fig. 3.6), there exists one particularly important structural difference between these two kinds of systems. Mainly, a true link-based system is necessarily connected as each new elements creates at least one link with the old elements, while node-based systems can create many disconnected components that may or may not end up interconnecting through new structures (depending on $q$ and $p$). In real link-based systems, there is no restriction on connectedness. For instance, the *cond-mat arXiv* network of co-authors has one giant component which consists of $\sim 93\%$ of the system, but other smaller satellite components still exist. While both SPA versions illustrated on Fig. 3.6 create a similar community structure as the *cond-mat arXiv*, the node-based version is actually closer to reality.

### 3.6.2 Multiple memberships, multiple links and self-loops

In our description of the time evolution of SPA, we have never explicitly forbidden an element to join the same structure more than once. These multiple memberships, whose like-

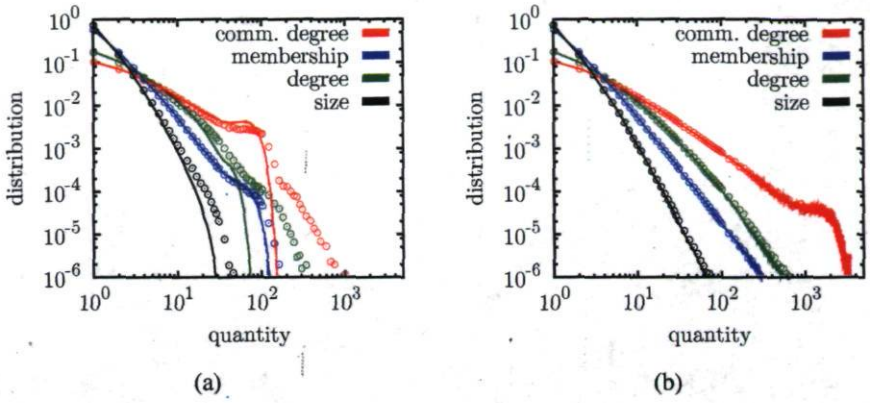



FIG. *3.7:* **Finite size effects in SPA.** Comparison between the time evolution data presented in Fig. 3.2 (dots) and the same data when multiple memberships, multiple links and self-loops are discarded (lines) for systems with a) 250 structures and b) 25 000 structures. Multiple memberships, multiple links and self-loops are finite size effects whose importance becomes null in the large-size limit.

lihood depends directly on the value of the $p$ or $q$ parameters, lead to multiple links between the same individuals and self-loops (where an element shares a structure with itself). Similarly, in our derivation of the degree distributions, we have supposed an infinite system where the probabilities that two structures overlap by more than one element fall to zero.

In empirical data, multiple links and self-loop are rarely considered. It can thus be useful to have an idea of the effect of such restrictions on SPA. Fig. 3.7 presents two snapshots of the same scenarios of SPA, with or without forbidding multiple memberships, multiple links and self-loops when analyzing the final stage of the system. The cutoffs in the distributions of the first system are not surprising, as large and old structures are very likely to have recruited the same element more than once, especially with a small $q$. Yet, this effect rapidly becomes negligible as the system grows and we enter the large size limit in accordance with the assumptions of our analytical description (see Fig. 3.7(b)).

### 3.6.3 Element-structure correlations

Most of the approximations used throughout this paper are based on the hypothesis of homogeneous mixing : the elements belonging to a number $x$ of structures *see* the same size distribution as the elements belonging to $y$ structures. This implies that there is no correlations except for the fact that an element is ten times more likely to belong to a given structure of size ten than to a particular structure of size one (*natural correlations*). To investigate this matter, we compare the size distributions as seen from elements with different memberships in both the simulations done for Fig. 3.6 and the corresponding *arXiv* data.

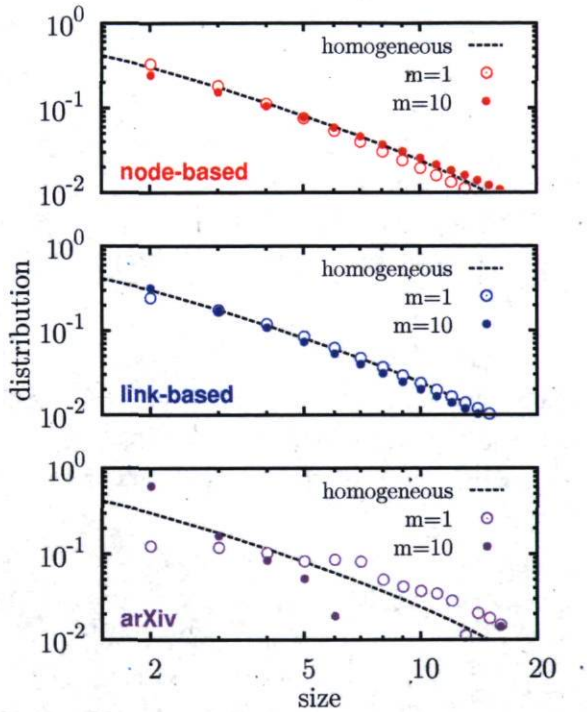



FIG. *3.8:* **Element-structure correlations in SPA and in the arXiv.** Size distribution of structures as seen from elements with different $m$ memberships. Markers represent empirical measures done on the *cond-mat arXiv* and numerical results on the two SPA processes (using the parameters of Fig. 3.6). The dotted line corresponds to what would be obtained through homogeneous pairing of memberships and structures. The small difference between the node-based and link-based SPA processes is most likely due to the fact that the link-based version requires more elements for the birth of new structures, which are consequently more likely to be old elements than in the node-based version.

Fig. 3.8 presents the results of this investigation. First, the similitude between SPA and homogeneous mixing explains why our approximations were accurate. Second, there is a major difference between element-structure correlations in real-systems and SPA : elements with few memberships are much more likely to belong to larger structures in the arXiv data than in our SPA simulations. This shows how other levels of organization have yet to be taken into account in our stochastic models. Depending on what one wants to model, these correlations could potentially be important.

## 3.7 Peloton dynamics

One particularly interesting feature of the results presented in Fig. 3.2 and 3.3 is the dynamics of the entities which were in the tail of the distributions. In fact, these groups of individuals or structures resulted in clearly identifiable *bulges* on their respective distributions. The dynamics of a system's leader is well-documented in the context of growing networks [41; 26] or word frequencies [11], but can be applied to any problem where one is interested in the statistics of the extremes (i.e. the growth of the biggest business firm, of the most popular website, etc.). What we observe here is that averaging over multiple realizations of the same experiment will result in the creation of a *peloton* where one is significantly more likely to find entities than predicted by the asymptotic distribution (i.e. the leaders).

The clear distinction between the statistical distribution of leaders versus the rest of the system is a consequence of the maximal size of the system and of the limited growth resources available. To illustrate this claim, we can consider a continuous time version of PA in which there is no finite limitation to the number of growth events at every time step (see Appendix for explicit solution of this process). Comparing the results of the discrete and continuous versions of our stochastic process on Fig. 3.9(a) illustrates how limiting growth resources results in the condensation of the leaders in a peloton. This draws a strong parallel between discrete preferential attachment and some sandpile models known to result in scale-free avalanche size distributions through *self-organized criticality*. In some cases, such as the Oslo model (see [12] §3.9), the biggest avalanches are limited by the size of the considered sandpile and are thus condensed in bulges identical to our pelotons.

Also striking is the fact that this peloton conserves its shape on a log-log scale (see Fig. 3.9(b)). To highlight this feature, Fig. 3.9(c) rescales the distributions to account for the scaling in size ($\gamma_s$) and the peloton growth through time ($t^{1-p}$). This rescaling method was borrowed from [12] §3.9.8.

Leaders emerge in every single preferential growth realization, while the peloton dynamics can only manifest itself once we average over multiple systems or over many characte-

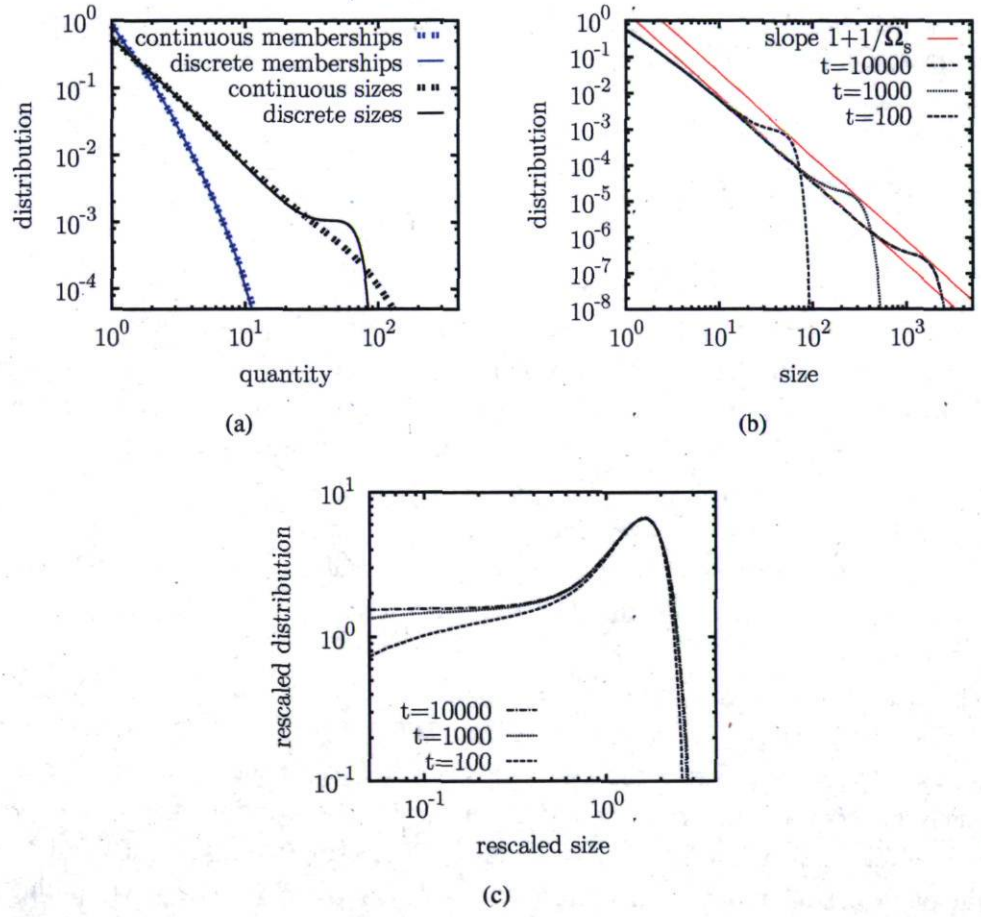



FIG. *3.9:* **Theoretical observation of peloton dynamics.** (a) Comparison of the memberships and sizes distributions of node-based SPA with $q = 0.8$ and $p = 0.2$ in discrete and continuous dynamics at time $t = 100$. This illustrates how the peloton dynamics is a direct consequence of the maximal system size present only in the discrete version of the process. (b) The height of the peloton follows a power-law decay (here for the results of Fig. 3.3(b)) , such that its surface is conserved on a logarithmic scale as it evolves. The decay exponent of the peloton is the same as the scaling exponent of the distribution it creates. (c) Rescaled distribution $\{n^{\gamma_s} S_n(t)\}$ as a function of rescaled community size $n/t^{1-p}$ to highlight the scaling of the peloton dynamics.

ristic time scales of a single system (through the births and deaths of many different leaders). Consequently, empirical observations of this phenomenon are rare, because on the one hand we have only one Internet, one arXiv, and basically a unique copy of most complex systems, and on the other hand, we have rarely access to extensive data through long time scales. We can however find a solution if we go back to the first example used by Simon [72] to derive his model : the scale-free distribution of words by their number of occurrences in written text (i.e. Zipf's law [84]). In this context, $q$ equals zero and the $p$ parameter corresponds to the probability that each new written word has never been used before. We can therefore consider different samples of text of equal length written by the same author as different realizations of the same experiment.

With this in mind, we have picked different authors according to personal preferences and size of their body of work and divided their oeuvres in samples of given lengths which we have used to evaluate Zipf's law under averaging (see Fig. 3.10). As predicted by PA, taking the average of multiple realizations of the same experiment results in a peloton which diverges from the traditional Zipf's law. In this case, the peloton implies that the leaders of this system (i.e. the most frequent words) consistently fall in the same scale of occurrences.

Lastly, Fig. 3.10(d) reproduces the scaling analysis of Fig. 3.9(c) for empirical results on prose samples. The varying surface of the peloton hints at a non-constant growth rate : a well-known feature of written text (see [28] §7.5).

## 3.8 Conclusion

In this paper, several analytical results for *structural preferential attachment* have been obtained : solutions of its time evolution, its asymptotic behavior and approximations for its different degree distributions. Those approximate descriptions are especially useful when it comes to using organization models as part of modelling efforts.

We have also highlighted one particular shortcoming of the model : element-structure correlations. That is, SPA lacks any modelling or predictive power when it comes to asking *who belongs to what structure.*

On the other hand, we have observed an interesting behavior of both the SPA and the classic PA models : *the peloton dynamics*. This particular feature is important in order to predict the position of the leaders of a PA growth process. More interestingly, we have been able to observe this behavior in the growth of prose samples, which differentiates the PA principle from the other models generating scale-free designs but failing to predict this property.

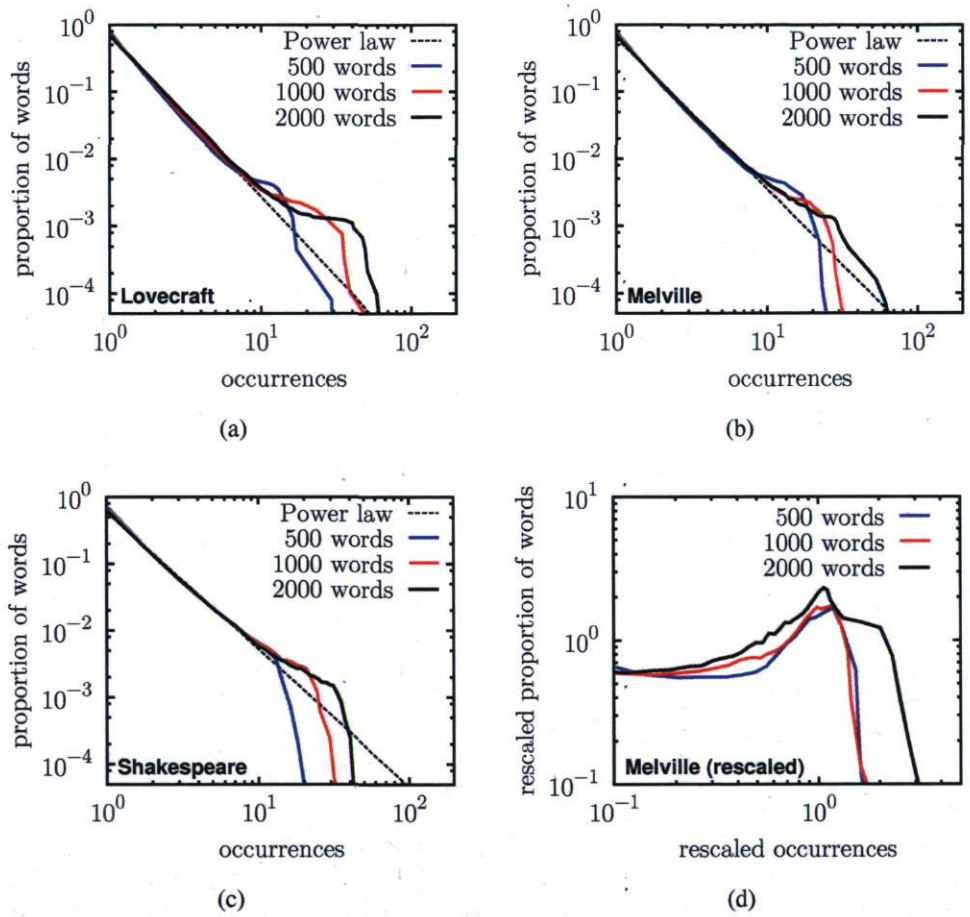



FIG. *3.10:* **Peloton dynamics in prose samples.** Distributions of words by their number of occurrences in prose samples of different length taken from the complete works of (a) H.P. Lovecraft composed of nearly 800 000 words, (b) William Shakespeare with around 900 000 words and (c) Herman Melville with over 1 200 000 words. The peloton dynamics is manifest in all distributions. (d) The rescaling method of Fig. 3.9(c) is here applied to the statistics of Herman Melville's work.

The presentation of shortcomings and successes of the SPA principle (in terms of predictive value) shows the importance and the need for further study in stochastic growth models.

## 3.9 Appendix : Explicit solution to continuous time SPA

Section 3.7 has presented an explicit solution for the time evolution of SPA in continuous time. This Appendix summarizes its derivation, based on a recently proposed method [49].

### 3.9.1 Definition of a continuous time PA process

The transition to continuous time simply implies that $q$ and $p$ will now refer to birth rates for both elements and structures. The corresponding rates $1-q$ and $1-p$ will thereby correspond to the growth rates of existing elements and structures, respectively. This means that in a given time interval $[t, t+1]$, this new stochastic process could create an infinite number of elements with probability $\lim_{dt\to 0}(qdt)^{1/dt}$ ; whereas the discrete version could only create one element with probability $q$. While it is highly improbable that continuous time PA results in a system several orders of magnitude larger than $qt$ or $pt$, there is no maximal size per se.

This sort of continuous time dynamics is better described using simple ODEs, or master equations as was done in [29]. To this end, we will once again follow $\tilde{N}_m$, the number of elements with $m$ memberships, and $\tilde{S}_n$, the number of structures enclosing $n$ elements. Using the same logic behind equations (3.1) and (3.2), but considering infinitesimal time steps $dt$, one can write

$$\tilde{N}_m(t+dt) = \tilde{N}_m(t) + dt\left\{\frac{\Gamma_s}{t}\left((m-1)\tilde{N}_{m-1}(t) - m\tilde{N}_m(t)\right) + q\,\delta_{m,1}\right\}$$

and

$$\tilde{S}_n(t+dt) = \tilde{S}_n(t) + dt\left\{\frac{\Omega_s}{t}\left((n-1)\tilde{S}_{n-1}(t) - n\tilde{S}_n(t)\right) + p\,\delta_{n,s}\right\},$$

which are straightforwardly rewritten as two ODEs :

$$\frac{d}{dt}\tilde{N}_m(t) = \frac{\Gamma_s}{t}\left((m-1)\tilde{N}_{m-1}(t) - m\tilde{N}_m(t)\right) + q\,\delta_{m,1}\ ; \tag{3.25}$$

$$\frac{d}{dt}\tilde{S}_n(t) = \frac{\Omega_s}{t}\left((n-1)\tilde{S}_{n-1}(t) - n\tilde{S}_n(t)\right) + p\,\delta_{n,s}\ . \tag{3.26}$$

Because these two last equations have the same form, we will solve them separately using a general continuous time PA equation. Consider

$$\frac{d}{dt}P_k(t) = \beta\,\delta_{k,m} + R_{k-1}(t)P_{k-1}(t) - R_k(t)P_k(t) \tag{3.27}$$

where $\beta$ is the birth rate, $m$ is the size of new entities and $R_i(t)$ is the attachment rate on entities of size $i$, which we define using a growth rate $\alpha$, an initial total size $m_0$ and a normalization rate $\lambda$ :

$$R_i(t) = \frac{\alpha i}{m_0 + \lambda t} . \tag{3.28}$$

It will prove useful to rewrite (3.27) in dimensionless form as

$$\frac{d}{d\tau} P_k(\tau) = \overline{\beta}\,\delta_{k,m} + \overline{R}_{k-1}(\tau) P_{k-1}(\tau) - \overline{R}_k(\tau) P_k(\tau) \tag{3.29}$$

with dimensionless time $\tau = \alpha t$, parameters $\overline{\beta} = \beta/\alpha$, $\overline{\lambda} = \lambda/\alpha$, and attachment rate $\overline{R}_k(\tau) = k/(m_0 + \overline{\lambda}\tau)$ respectively. Table 3.1 gives the values of the different parameters for the classical PA models and for SPA.

| | PA | | SPA | |
|---|---|---|---|---|
| | Simon | BA | elements | structures |
| $\beta/\alpha$ | $p/(1-p)$ | $1/m$ | $q/\alpha$ | $p/\alpha$ |
| $\alpha$ | $1-p$ | $m$ | $1-q+p(s-1)$ | $1-p$ |
| $\lambda/\alpha$ | $1/(1-p)$ | 2 | $[1+p(s-.1)]/\alpha$ | $[1+p(s-1)]/\alpha$ |
| $m$ | 1 | $m$ | 1 | $s$ |

TAB. *3.1:* **Parameters of a general preferential attachment process.** Parameters of the process described by (Eq. 3.29) in the context of Simon's model [72], of the Barabási-Albert model (BA) [9] and of SPA.

### 3.9.2 Explicit solution

Let

$$\overline{H}_k(t) = \exp\left[\int \overline{R}_k(\tau) d\tau\right] = \left(m_0 + \overline{\lambda}\tau\right)^{k/\overline{\lambda}} , \tag{3.30}$$

so that Eq. (3.29) can be written as :

$$\frac{d}{d\tau}\left[P_k(\tau)\overline{H}_k(\tau)\right] = \overline{\beta}\overline{H}_k(\tau)\delta_{k,m} + \overline{R}_{k-1}(\tau)\overline{H}_k(\tau)P_{k-1}(\tau) . \tag{3.31}$$

The general solution of this transformed equation is :

$$P_k(\tau) = \overline{\beta}\frac{(m_0 + \overline{\lambda}\tau)}{k + \overline{\lambda}}\delta_{k,m} + \frac{(1-\delta_{k,m})}{\overline{H}_k(\tau)} \int \overline{R}_{k-1}(\tau)\overline{H}_k(\tau)P_{k-1}(\tau)d\tau + C_k , \tag{3.32}$$

where $\{C_k\}$ are constants of integration determined by the initial conditions. Solving for the first few values of $k$ ($m$, $m+1$, $m+2$, ...) reveals the following pattern for the solutions :

$$P_{m+k}(\tau) = \overline{\beta}\frac{(m)_k}{(m+\overline{\lambda})_{k+1}}\left(m_0 + \overline{\lambda}\tau\right) + \sum_{i=0}^{k}\frac{(m)_k}{(m)_i}\frac{C_{m+i}}{(k-i)!}\left(m_0 + \overline{\lambda}\tau\right)^{-(m+i)/\overline{\lambda}} \tag{3.33}$$

where $(\gamma)_j \equiv (\gamma)(\gamma+1)\ldots(\gamma+j-1)$ are Pochammer symbols. The last step towards a complete solution is to determine an explicit form of the constants of integrations $\{C_{m+k}\}$ in terms of the initial conditions $\{P_{m+k}(0)\}$. This is easily accomplished by writing (3.33) in a matrix form for the vector of initial conditions $\boldsymbol{P}(0)$

$$\boldsymbol{P}(0) = \boldsymbol{A}(0) + \mathbf{L}(\mathbf{0})\boldsymbol{C} \tag{3.34}$$

in terms of the vector $\boldsymbol{C}$ of integration constants and a *lower triangular* matrix $\mathbf{L}$, followed by the observation that the inverse of a (lower/upper) triangular matrix is also a (lower/upper) triangular matrix whose elements can be constructed by forward substitution. Given that the elements of $\mathbf{L}(0)$ are

$$L_{m+k,m+i}(0) = \binom{m+k-1}{m+i-1}\frac{1}{m_0^{m+i}} \tag{3.35}$$

we find that the elements of the inverse matrix, denoted $\mathbf{M}$, are simply

$$M_{m+k,m+i} = (-1)^{k-i}\binom{m+k-1}{m+i-1}m_0^{m+i}\,. \tag{3.36}$$

Inserting this solution in (3.33), we get

$$\boldsymbol{P}(\tau) = [\boldsymbol{A}(\tau) - \mathbf{L}(\tau)\mathbf{M}\boldsymbol{A}(0)] + \mathbf{L}(\tau)\mathbf{M}\boldsymbol{P}(\mathbf{0})\,, \tag{3.37}$$

which nicely isolates the principal dynamics (the first 2 terms) from the initial conditions. Specifically, by imposing the usual initial conditions, $P_{m+k}(0) = \delta_{k,0}$, it is straightforward, albeit somewhat lengthy, to obtain a closed-form expression for the complete dynamical elements as

$$\begin{aligned} P_{m+k}(\tau) &= \overline{\beta} m_0 (m)_k \left[\frac{1}{(m+\overline{\lambda})_{k+1}} X(\tau) - \frac{1}{(m+\overline{\lambda})}\frac{1}{\Gamma(k+1)} X(\tau)^m F_k(X(\tau))\right] \\ &+ (m)_k \frac{1}{\Gamma(k+1)} X(\tau)^m (1-X(\tau))^k \end{aligned} \tag{3.38}$$

with $X(\tau) = m_0/(m_0+\overline{\lambda}\tau)$ and where $F_k(X) = {}_2F_1(-k, m+\overline{\lambda}; m+\overline{\lambda}+1; X)$ represents a terminating hypergeometric series of degree $k$. One verifies that, by setting $\tau = 0$ in the previous expression, one obtains $P_{m+k}(0) = \delta_{k,0}$ as it should.

It can further be shown that the continuous and discrete time versions of PA converge toward the same asymptotic behavior.

# Chapitre 4

# Dynamique de propagation sur structure communautaire

**Propagation dynamics on networks featuring complex topologies**

Laurent Hébert-Dufresne, Pierre-André Noël, Vincent Marceau,
Antoine Allard et Louis J. Dubé
Département de Physique, de Génie Physique, et d'Optique,
Université Laval, Québec, Québec, Canada G1V 0A6.

Référence : Physical Review E, 82 (2010), p. 036115.

© 2010 The American Physical Society

## 4.1 Avant-propos

Les chapitres précédents s'intéressaient à la modélisation de la topologie des réseaux complexes et plus précisément à la description et à l'émergence de leur structure communautaire. Ici, on met en lumière l'influence de cette structure communautaire sur une application de haute importance sociétale : la dynamique propagatoire sur réseaux complexes. On développe d'abord une description plus générale de la structure communautaire que ce que nous avons employé jusqu'ici, dans le but que le formalisme analytique développé soit indépendant de l'attachement préférentiel structurel.

## 4.2 Résumé

La description analytique de processus propagatoires sur réseaux aléatoires a connu un véritable essor dans les dernières années, mais les systèmes plus complexes ont principalement été étudiés numériquement. Dans cet article, une description par champs moyens est utilisée pour coupler de façon cohérente la dynamique des éléments du réseau (noeuds, individus...) d'une part et des motifs récurrents de leur topologie (structures, communautés...) d'autre part. Dans un modèle SIS de propagation d'épidémie sur réseaux sociaux à structure communautaire, cette approche procure un système d'EDO pour l'évolution temporelle du système, en plus de solutions analytiques pour le seuil épidémique et l'état endémique d'équilibre. Les résultats obtenus sont en bon accord avec les simulations numériques et reproduisent le comportement de réseaux aléatoires dans les limites appropriées, ce qui aide à illustrer l'influence de la topologie sur la dynamique. Finalement, il est démontré que, en absence de corrélation en degrés, notre modèle prédit un seuil épidémique plus élevé pour une structure communautaire que pour une topologie aléatoire équivalente.

## 4.3 Abstract

Analytical description of propagation phenomena on random networks has flourished in recent years, yet more complex systems have mainly been studied through numerical means. In this paper, a mean-field description is used to coherently couple the dynamics of the network elements (nodes, vertices, individuals...) on the one hand and their recurrent topological patterns (subgraphs, groups...) on the other hand. In a SIS model of epidemic spread on social networks with community structure, this approach yields a set of ODEs for the time evolution of the system, as well as analytical solutions for the epidemic threshold and equilibria. The results obtained are in good agreement with numerical simulations and reproduce random networks behavior in the appropriate limits which highlights the influence of topology on

the processes. Finally, it is demonstrated that, in the absence of degree correlation, our model predicts higher epidemic thresholds for clustered structures than for equivalent random topologies.

## 4.4 Introduction

Description of propagation phenomena has been one of the most prolific field in complex network theory, mostly because of the range of possible applications : epidemic control, spread of information, virus or pollutant propagation in electronic or biological networks [10]. Most analytical models are based on the random network (RN) paradigm : from the point of view of the propagating agent, random networks are seen as identical for every newly infected individual because of their treelike structure (i.e. no loops). This approach has given rise to different descriptions : some are based on a compartmentalisation of nodes according to their state [5], others on the generating function formalism [50; 51; 4; 59] or hybrid descriptions using mean-field theory [78; 46] ; yet all approaches are difficult to generalize to real networks for which the RN paradigm rarely applies.

The importance of topology for propagation dynamics [66; 78; 47; 37; 38; 71; 51], and more specifically, the importance of clustering [48; 24; 56; 19; 52; 58], is now well established. That is, the dynamics on the network is far from independent on how links are arranged between its elements. Furthermore, most real networks feature a significant amount of substructures that simply cannot be ignored as they define the very identity of the networks. The multi-protein units of molecular biology [69; 76], the coupling of a given set of stocks [60; 31] or groups of highly connected individuals [61; 58] are all good examples of how precise mechanisms (e.g. the friend of my friend is my friend) give rise to important structures within a seemingly random topology.

The two limits of complex networks, complete randomness and perfect order, can be treated with the previously discussed methods. We will concentrate on those particular complex networks, located somewhere between order and disorder, and show how their topology can be taken into account in dynamical problems. In doing so, the language of social networks and epidemics will be used to take advantage of its eloquence and clarity. It should be clear however that the formalism developed is general to many types of networks and propagation phenomena.

The paper is structured as follows. The particular topology chosen to illustrate our approach, the community structure (CS), is described in Sec. 4.5. The analytical model is then developed in Sec. 4.6 where we also obtain analytical solutions for the equilibria and epidemic threshold of the system. Section 4.7 compares our analytical results with numerical

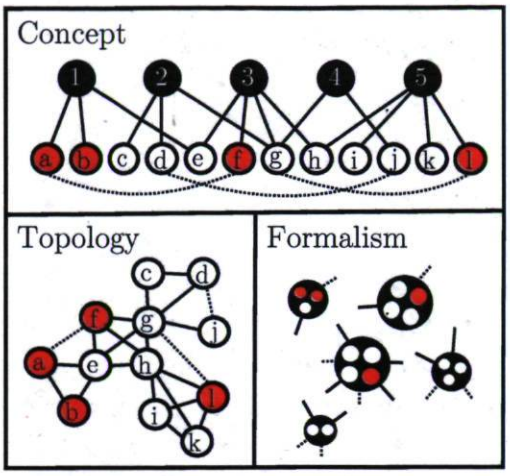



FIG. *4.1:* **Community structure : topology and formalism.** Schematization of the particular topology studied in this paper. An open mark represents a susceptible individual ; a red one, an infectious ; and a black one, a group (or clique). The topology is constructed by allowing individuals to belong to a given number of cliques where they can be linked to other participants (solid lines) and then randomly assigning random exterior neighbors (dotted lines). Note that in the formalism, the cliques are differentiable by their exact population and state, while the precise connections between them remain unspecified and they are simply linked to a mean-field.

Monte-Carlo (MC) simulations and presents discussions of our findings. After presenting our conclusions in Sec. 4.8, an Appendix completes our analysis of propagation phenomena on community structure.

## 4.5 Community structure

In what follows, a new approach to describe dynamical problems on complex topologies will be used to solve a disease propagation model on social networks featuring a well-known topology : the community structure. We define this particular arrangement of nodes by their aggregation in highly connected groups. These communities (or cliques) can virtually represent a person's family, workplace, collection of friends, etc. This simple concept results in a network with highly connected communities and a sparser density of links between them (see Fig. 4.1). The topology of such networks has been studied at some length : for its initial description, see [23] ; for its statistical significance, see [35] ; for its detection or characterisation, see [55; 68; 57; 53; 14] ; and the references therein for an exhaustive presentation.

Unfortunately, not unlike other complex types of networks, studies of dynamical processes on this topology has been mainly limited to numerical simulations (e.g. [34]). Albeit

useful to estimate its effect on the dynamics, they lack the clarity of an analytical framework. On the other hand, mean-field description of propagation phenomena in terms of communities (or households) has been previously attempted in [6; 21; 33] with several shortcomings such as homogeneous topology, lack of the concept of individuals or inefficient moment closure approximations. Hence, there is a need for an analytical approach that can accurately take into account the many complexities of social networks in order to describe the time evolution of the system. Because community structure typically includes clustering and degree correlation, our formalism will include the coherent contribution of both properties.

A useful model of social topology was published by Newman in [52]. The networks are constructed as follows : each individual belongs to $m$ cliques and each clique holds $n$ individuals, where both $m$ and $n$ are taken from given distributions. Within every clique, each pair of members has a probability $\epsilon$ of being acquainted. Hence, the entire topology is defined by one parameter $\epsilon$ and two probability distributions $\{g_m\}$ and $\{p_n\}$ generated by the following probability generating functions (PGFs) :

$$P_0(z) = \sum_{n=0}^{\infty} p_n z^n , \tag{4.1}$$

$$G_0(z) = \sum_{m=0}^{\infty} g_m z^m . \tag{4.2}$$

which are simply built from the probabilities $p_n$ and $g_m$ that a random clique or individual will have $n$ participants or $m$ cliques respectively. Similar functions can be defined to generate the probabilities that a random clique of a random individual is shared by $n-1$ other participants or that a random individual in a random clique participates in $m-1$ other cliques. We simply note that these quantities are proportional to $np_n$ or $mg_m$ and thus find our second set of PGFs :

$$P_1(z) = \frac{\sum_n n p_n z^{n-1}}{\sum_n n p_n} = \frac{P_0'(z)}{P_0'(1)} = \nu^{-1} P_0'(z) , \tag{4.3}$$

$$G_1(z) = \frac{\sum_m m g_m z^{m-1}}{\sum_m m g_m} = \frac{G_0'(z)}{G_0'(1)} = \mu^{-1} G_0'(z) \tag{4.4}$$

where $\nu$ and $\mu$ are respectively the mean numbers of individuals per clique and cliques per individual used to normalize the distributions. Note that the mean of a distributed quantity is simply given by the derivative of the corresponding PGF evaluated at unity. The following topological properties have already been derived in [52] and [58] : degree distribution, size of the giant component, clustering coefficient and degree correlation. Some of these results are used throughout this paper.

Newman's model, although realistic because of its overlapping communities, is strongly limited since links only arise through communities. A node belonging to a single clique does not participate at all in the coupling, while a node belonging to two cliques or more will have

a huge influence. Hence, it is hard to describe weakly coupled communities of significant sizes using this particular topology. Consequently, we will introduce a more general description of community structure where exterior random links are also allowed. We simply add a distribution for the number of random links per individual, which is generated by :

$$K_0(z) = \sum_{l=0}^{\infty} k_l z^l \,. \tag{4.5}$$

Our networks will thus be defined by the $\epsilon$ probability and three distributions for the numbers of individuals per clique (4.1), cliques per individual (4.2) and random links per individual (4.5). Intuition indicates that a large number of networks can be decomposed as basic structures coupled either by sharing nodes, by forced connections or a combination of both. In fact, many of the previously cited papers study networks where nodes belong to a single clique coupled only by random links with the outside world (e.g. [57; 24]). Our general model includes this topology and Newman's original model as special cases.

## 4.6 SIS model of epidemics on community structure

### 4.6.1 Construction of the dynamical model

The philosophy behind our formalism is to analyze the network simultaneously from two perspectives, i.e. the state of the network is followed from the point of view of recurrent patterns in its topology and of the elements themselves. More precisely, we compartmentalize both the structure and the node ensemble in terms of their relation to one another and couple the two systems to give a complete description of the propagation phenomenon. For social networks featuring community structure, the recurrent patterns are cliques of individuals that can be distinguished by their size and their state. The elements are individuals distinguishable by the number of cliques to which they belong and by their number of exterior random links. That is, the mean state of a given class of individuals will act as if all of their cliques and random links were approximated by a mean-field and the mean state of a given class of cliques will act as if all individuals were also reduced to a mean-field approximation. The behaviors of both cliques and individuals are coupled in terms of their connections via the generating functions (4.1) through (4.5).

The particular case under study is a Susceptible-Infectious-Susceptible (SIS) model of disease propagation. In continuous time, an infectious node may pass the disease to any of its susceptible neighbors at a rate $\tau$ (S $\rightarrow$ I), while it is recovering from the disease at a rate $r$ (I $\rightarrow$ S). Given initial conditions, we are interested in developing a system of equations capable of following the state $I(t)$ of the network, where $I(t)$ is the fraction of infectious individuals at a given time. According to our philosophy, we thus need to follow both individuals

and cliques. Let $S_{m,l}(t)$ be the proportion of individuals which belong to $m$ cliques, have $l$ random links and are susceptible at time $t$ and $C_{n,i}(t)$ be the proportion of cliques whose population is $n$ and of which $i$ are infectious at time $t$. For the sake of clarity, we will not explicitly mark the time dependence, $(t)$, when it is obvious that the quantity is a dynamical variable.

First, we need to describe how the generating functions $G_1(z)$, $K_0(z)$ and $P_1(z)$ will differ depending on the state of the involved individual. To define the dynamical generating functions, it is possible to either follow the distributions for the susceptibles or the infectious individuals, since $S_{m,l} + I_{m,l} = g_m k_l$. We will follow the susceptibles. We then need the distribution of cliques reached from a susceptible individual of a given clique. This distribution will be affected by $\{S_{m,l}\}$ in the following manner : a random individual has probability $mg_m$ of belonging to $(m-1)$ other cliques, but consequently, only a probability $\sum_l S_{m,l}/g_m$ of being susceptible at time $t$. The reasoning is even simpler for $K_0(z)$ as the distribution is not affected by the knowledge that the individual belongs to at least one clique. We can directly write :

$$\widetilde{G}_1(z;t) = \frac{\sum_{m,l} mS_{m,l} z^{m-1}}{\sum_{m,l} mS_{m,l}}, \tag{4.6}$$

$$\widetilde{K}_0(z;t) = \frac{\sum_{m,l} S_{m,l} z^l}{\sum_{m,l} S_{m,l}}, \tag{4.7}$$

where the tilde denotes that the function generates a distribution which applies to susceptible individuals only. In a similar fashion, the knowledge that a clique is reached by a link emerging of a susceptible individual will affect the distribution of this clique's number of susceptible individuals. The probability that a susceptible individual belongs to a clique of state $\{n, i\}$ is directly proportional to the number of susceptible members of that particular state. In order to consider only susceptibles individuals, the $P_1(x, y)$ generating function must be modified accordingly to the number of susceptible members belonging to each compartment :

$$\widetilde{P}_1(x, y; t) = \frac{\sum_{n,i}(n-i)C_{n,i}x^n y^i}{\sum_{n,i}(n-i)C_{n,i}}. \tag{4.8}$$

Four interesting and important quantities can be derived from these dynamical generating functions. Firstly, the average number of infectious neighbors per clique and per random link for a susceptible individual, $R(t)$ and $T(t)$ :

$$R(t) = \epsilon \frac{\sum_{n,i} i(n-i)C_{n,i}}{\sum_{n,i}(n-i)C_{n,i}}, \tag{4.9}$$

$$T(t) = \frac{\sum_{n,i} \frac{i}{n}(nC_{n,i})}{\sum_{n,i} nC_{n,i}}. \tag{4.10}$$

Secondly, the mean number of excess infectious neighbors per clique and per random link for a susceptible individual of a given clique, $\rho(t)$ and $\sigma(t)$ :

$$\rho(t) = \widetilde{G}_1'(1;t)R(t), \tag{4.11}$$

$$\sigma(t) = \widetilde{K}_0'(1;t)T(t) \tag{4.12}$$

where the primes denote a derivative with respect to $z$, so that $\widetilde{G}_1'(1;t)$ is the average number of outside cliques for a susceptible member of a given clique at time $t$.

Let us now construct the differential equation governing $\{S_{m,l}\}$. We previously mentionned that the disease spreads through any link between a susceptible and an infectious individual. Thus, with $R(t)$ being the average number of such links that a susceptible may have in a single clique, the rate at which the class of individuals belonging to $m$ cliques is infected, is proportional to $-\tau m S_{m,l} R(t)$. Similarly, with $T(t)$ being the probability that a random link leads to an infectious individual, the rate of infection for individuals with $l$ random links must be proportional to $-\tau l S_{m,l} T(t)$. Simultaneously, the same ratio increases as the infected nodes recover at a speed $r(g_m k_l - S_{m,l})$. Therefore, the set of equations governing the point of view of the individuals is simply obtained by summing the contributions from these three processes :

$$\frac{dS_{m,l}}{dt} = r(g_m k_l - S_{m,l}) - \tau S_{m,l}\left[mR(t) + lT(t)\right] . \tag{4.13}$$

Similar considerations are needed to define the dynamics of the $C_{n,i}$ values. A clique in a $\{n,i\}$ state can either pass to $\{n,i+1\}$ by infection (if $i < n$) or to $\{n,i-1\}$ by recovery (if $i > 0$). The first process is proportional to the sum of the number of links between infectious and susceptible individuals within the cliques and the number of links with infectious neighbors that each susceptible might have outside the considered clique. For a given $\{n,i\}$ compartment, infection can either bring new cliques from the $\{n,i-1\}$ state or cause the cliques to pass to the more infectious $\{n,i+1\}$ compartment :

$$\frac{dC_{n,i}}{dt} \propto \tau(n-i+1)\left[\epsilon(i-1) + \rho(t) + \sigma(t)\right] C_{n,i-1} - \tau(n-i)\left[\epsilon i + \rho(t) + \sigma(t)\right] C_{n,i} . \tag{4.14}$$

The contribution of the recovery process is easy to explicit using the same logic, as it is simply proportional to the number of infectious individuals who might recover :

$$\frac{dC_{n,i}}{dt} \propto r(i+1) C_{n,i+1} - r i C_{n,i} . \tag{4.15}$$

Summing the contributions of both the infections (4.14) and the recoveries (4.15) yields the desired differential equation for the cliques dynamics :

$$\begin{aligned}\frac{dC_{n,i}}{dt} = {} & r(i+1) C_{n,i+1} - r i C_{n,i} - \tau(n-i)\left[\epsilon i + \rho(t) + \sigma(t)\right] C_{n,i} . \\ & + \tau(n-i+1)\left[\epsilon(i-1) + \rho(t) + \sigma(t)\right] C_{n,i-1} .\end{aligned} \tag{4.16}$$

where $C_{n,i}$ is defined only for $i \in [0,n]$. Coupled with Eq. (4.13), we now have a complete dynamical system for the state of the network in a SIS model of disease spread.

If desired, the mean fraction of infectious individuals of a given class of cliques can be obtained in a straightforward manner with :

$$I_n = \sum_i \frac{1}{n p_n} i C_{n,i} . \tag{4.17}$$

It is generally simpler to characterize the state of the network via the total fraction of infectious, $I(t)$, or susceptible, $S(t)$, individuals. From Eq. (4.13), we directly have :

$$S(t) = \sum_{m,l} S_{m,l} \; ; \quad I(t) = \sum_{m,l} (1 - S_{m,l}) \; . \tag{4.18}$$

Note that a straightforward evaluation of the global state of the network from $\{C_{n,i}\}$ would be biased because an individual belonging to $m$ cliques would be counted $m$ times more than an individual participating to a single clique.

### 4.6.2 Solution for network stable state

System (4.13) and (4.16) can be solved as a traditional self-consistent field by looking for a solution in terms of $\rho$ and $\sigma$. Using Eq. (4.16) for the $C_{n,i}$ quantities at equilibrium (i.e. $dC_{n,i}/dt = 0$), we obtain the following recursive solution :

$$C^*_{n,i+1} = \frac{1}{(i+1)r} \left[ (f_{n,i} + ri) C^*_{n,i} - f_{n,i-1} C^*_{n,i-1} \right] \tag{4.19}$$

with $C_{n,i} = 0 \; \forall \; i \notin [0, n]$, and where we introduce a matrix of infection $\{f_{n,i}\}$ whose elements depend on the total mean-field $\xi$ :

$$f_{n,i} \equiv \tau(n-i)(i\epsilon + \xi^*) \tag{4.20}$$

$$\xi^* \equiv \rho^* + \sigma^* \; . \tag{4.21}$$

Asterisks will hereafter refer to values at equilibrium. Equation (4.19) can be used to fix the stable values of all the $C^*_{n,i}$ relative to $C^*_{n,0}$, which can then be solved exactly by applying the following topological constraint :

$$\sum_i C_{n,i} = p_n \quad \forall \; t, n \; . \tag{4.22}$$

Using the equilibrium condition on Eq. (4.13) provides a direct solution for the $S^*_{m,l}$ ensemble :

$$S^*_{m,l} = \frac{r}{\tau (mR^* + lT^*) + r} \; . \tag{4.23}$$

It is then possible to write $R^*$, $T^*$, $\widetilde{G}^*_1(z)$ and $\widetilde{K}^*_0(z)$ in terms of $\rho^*$ and $\sigma^*$ by using (4.19) in (4.9) and (4.10) while using (4.23) in (4.6) and (4.7). A transcendental equation is obtained for $\xi^*$ by writing (4.11) and (4.12) as :

$$\xi^* = \left[ \frac{\sum_{m,l} m(m-1) S^*_{m,l}}{\sum_{m,l} m S^*_{m,l}} \right] R^* + \left[ \frac{\sum_{m,l} l S^*_{m,l}}{\sum_{m,l} S^*_{m,l}} \right] T^* \equiv F(\xi^*) \; , \tag{4.24}$$

where the dependence on $\xi^*$ comes from that of $\{S^*_{m,l}\}$ on $R^*$ and $T^*$ written in terms of $\{C^*_{n,i}\}$ which are a direct function of $\xi^*$. Solving for $\xi^*$ yields a unique non-zero solution fixing $\{C^*_{n,i}\}$

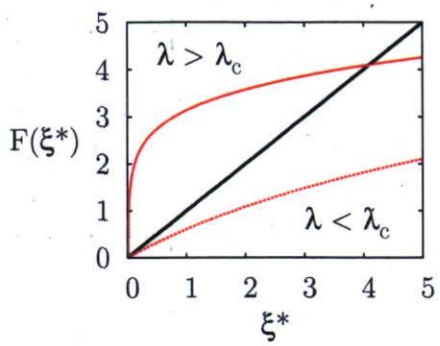



FIG. 4.2: **Identifying the epidemic threshold and equilibrium of a disease.** Function $F(\xi^*)$ is shown in red on the topology defined in (4.32) for two different normalized propagations rates : $\lambda = 0.02$ in dotted line (under the threshold; no solution for $\xi^* > 0$) and $\lambda = 0.1$ in solid line (epidemic). The black solid line is the curve of slope 1, $F(\xi^*) = \xi^*$.

which in turn provide the values for $R^*$ and $T^*$. This directly fixes $\{S^*_{m,l}\}$ using (4.23), and thus the stable state of the network defined by (4.18).

Clearly the dynamics is governed by the ratio $\lambda \equiv \tau/r$ and not the individual rates. Therefore, under the transformation to the normalized propagation rate $\lambda$, our model admits a single independent parameter in its dynamics.

### 4.6.3 Solution for epidemic threshold

The epidemic threshold $\lambda_c$ is defined by a phase transition in the normalized infection rate where a macroscopic final epidemic size first appears. Here, it can be defined mathematically using the analytic solution for the stable state of the SIS epidemic. Equation (4.24) behaves as shown in Fig. 4.2 with a trivial solution at $\xi^* = 0$ and another possible solution $\xi^* > 0$ depending on $\lambda$ and the topology. Since $F(\xi^*)$ is a monotonously increasing function, $\lambda_c$ can be found by the following condition :

$$\frac{d}{d\xi^*} F(\xi^*)\Big|_{\xi^*=0} = 1 \,. \tag{4.25}$$

For initial derivative value above unity, a solution $\xi^* > 0$ exists and the stable epidemic state is non-zero (Fig. 4.2). For a system subject to a propagation at its threshold, by definition, we know that the stable state is the trivial solution $C^*_{n,i} = p_n \delta_{i0} \; \forall \; \{n, i\}$ and $S^*_{m,l} = g_m k_l \; \forall \; \{m, l\}$ (which implies $\widetilde{G}_1(z;t) = G_1(z)$ and $\widetilde{K}_0(z;t) = K_0(z)$). It follows that the mean-field values are zero at equilibrium and (4.25) straightforwardly becomes :

$$\frac{1}{\nu} \sum_{n,i} \left\{ \epsilon i(n-i) G_1'(1) + i K_0'(1) \right\} \frac{d}{d\xi^*} C^*_{n,i}\Big|_{\xi^*=0} = 1 \,. \tag{4.26}$$

Using (4.19) to evaluate the derivative at equilibrium, one finds that $\forall\ i > 0$ :

$$\left.\frac{d}{d\xi^*}C^*_{n,i}\right|_{\xi^*=0} = \frac{p_n}{i}\lambda_c^i\epsilon^{i-1}\frac{n!}{(n-i)!}\ . \tag{4.27}$$

Using (4.27) to solve (4.26) for $\lambda_c$ provides a polynomial with positive coefficients for terms of order one or more :

$$\frac{1}{\nu}\sum_{n,i>0} p_n\,(\epsilon\lambda_c)^i\,\frac{n!}{(n-i)!}\left((n-i)G_1'(1) + \frac{K_0'(1)}{\epsilon}\right) = 1\ . \tag{4.28}$$

This polynomial therefore has a single real positive solution, which is the epidemic threshold of the network. For random networks, one can set $K_0'(1) = 0$, $\epsilon = 1$ and $p_n = \delta_{n,2}$ so that all links are shared within cliques of size two. Expression (4.28) then reduces to :

$$G_1'(1)\lambda_c^{\mathrm{RN}} = 1\ , \tag{4.29}$$

where $G_1'(1)$ is here the mean excess degree. From Eq. (4.29), one can deduce that our model predicts a null SIS epidemic threshold only if $G_1'(1)$ diverges. For scale-free networks whose degree distribution falls as $k^{-s}$, it can be shown that $G_1'(1)$ diverges if $s \leq 3$. Our model therefore leads to the same conclusion as [65] : scale-free networks with degree distribution $p_k \propto k^{-s}$ and $s \leq 3$ are defined by an absence of epidemic threshold.

A calculation of the SIS epidemic threshold on random networks was previously done in [64], using discrete time steps and constant recovery period approximations. To the best of our knowledge, Eq. (4.28) is the first equation for a continuous time SIS model of epidemic spread for both random networks and community structure.

## 4.7 Implementation and validation

### 4.7.1 Treatment of the analytical model

In order to highlight the difference between RN and CS, both types of networks will be studied analytically and numerically. The CS network will be compared with its equivalent random network (ERN) : a network with exactly the same degree distribution, but with randomly connected nodes (zero degree correlation). Note that on our general model of community structure, the PGF for the degree distribution is simply generated by [52] :

$$G_0\left(P_1(1 + (z-1)\epsilon)\right) \times K_0(z)\ . \tag{4.30}$$

To describe an ERN with this distribution, two simple options are available. Firstly, one can set $P_0^{\mathrm{ERN}}(z) = z^2$ and $K_0^{\mathrm{ERN}}(z) = 1$ with $\epsilon^{\mathrm{ERN}} = 1$ so that all cliques are of size two (i.e. regular

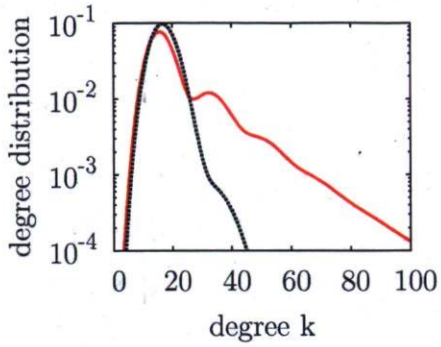



FIG. *4.3:* **Degree distributions of two community structured system.** Degree distribution in the infinite network limit of the chosen topologies : (4.33) is shown by a solid red line while (4.35) is shown by a dotted black line. Note the periodic local maxima corresponding to each $m$ value.

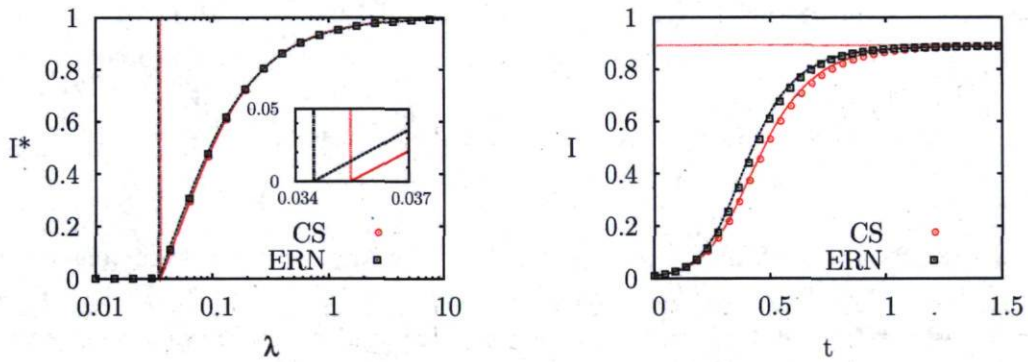



FIG. *4.4:* **Time evolution and steady state of an epidemic on community structure.** Comparisons of analytical and numerical results on a network defined by (4.32) using normalized dynamics ($t \to rt$ and $\lambda = \tau/r$). (a) analytical stable states (curves) and epidemic thresholds (vertical lines at $\lambda_c^{\mathrm{CS}} = 3.54 \cdot 10^{-2}$ and $\lambda_c^{\mathrm{ERN}} = 3.44 \cdot 10^{-2}$). (b) time evolution (curves) and analytical equilibrium (horizontal line) for $\lambda = 0.5$. On both figures, the results are shown in solid red for the community structure (CS) and in dotted black line for the equivalent random network (ERN). Numerical results are presented by markers and are averaged over 20 000 networks of 25 000 nodes. The standard deviation is smaller than the marker size.

links) and then choose the $g_m$ distribution equal to the initial degree distribution (4.30) of the CS network. Secondly, one can set $P_0^{\mathrm{ERN}}(z) = z$ and $G_0^{\mathrm{ERN}}(z) = z$ with any $\epsilon^{\mathrm{ERN}}$ so that all cliques are of size one (i.e. simple nodes) and then choose the $k_l$ distribution equal to the initial degree distribution (4.30). Both will be used in what follows.

The time evolution of the analytical system is obtained from an integration based on a 4th

order Runge-Kutta algorithm with adaptive time steps. The initial condition $I(0)$ is uniformly distributed among the nodes. That is, $S_{m,l}(0) = g_m k_l (1 - I(0))$ for all $\{m, l\}$, while $\{C_{n,i}(0)\}$ are given by a simple Bernoulli trial :

$$C_{n,i}(0) = p_n \binom{n}{i} [I(0)]^i [1 - I(0)]^{n-i} \ . \tag{4.31}$$

### 4.7.2 Numerical model

To perform MC simulations of the model, we have generated networks with the structure presented in section 4.5 via the following numerical algorithm :

i. generate a sequence $\{m_i\}$ of length $N$ subjected to distribution $\{g_m\}$ ;

ii. generate a sequence $\{n_j\}$ subjected to distribution $\{p_n\}$ until $\sum_j n_j = \sum_i m_i$ ;

iii. for each $i$, produce $m_i$ individuals tagged as $i$ ;

iv. for each $j$, produce $n_j$ groups tagged as $j$ ;

v. randomly assign each individual to a group ;

vi. for each $i$, list every $i$ assigned to the $n_j$ groups and link them to one another with probability $\epsilon$.

vii. generate a sequence $\{l_s\}$ of length $N$ subjected to the distribution $\{k_l\}$ under condition that $\sum_s l_s$ is even ;

viii. for each $s$, produce $l_s$ stubs tagged as $s$ ;

ix. randomly link all stubs in pairs.

The final ensemble of links presents a topology as shown in Fig. 4.1 with a degree distribution generated by (4.30) ; where nodes are highly clustered, but the clique concept itself is invisible. Each and every network generated by this procedure is accepted and kept in the results, as they are part of the canonical ensemble considered by the mean-field approach of the formalism. For every generated network, a fraction $I(0)$ of individuals are randomly chosen to be initially infectious and the dynamics is then simulated in a discrete time propagation simulation valid for a time step $\Delta t \to 0$ (we choose $\Delta t$ such that $\tau \Delta t$ and $r \Delta t$ are lesser than $10^{-3}$) :

i. at each $\Delta t$, every susceptible neighbor of every infectious individual is infected with probability $\tau \Delta t$ ;

ii. at each $\Delta t$ every infectious individual recovers with probability $r \Delta t$.

Finally, for each constructed network, the final degree distribution is used to generate an ERN for comparison.

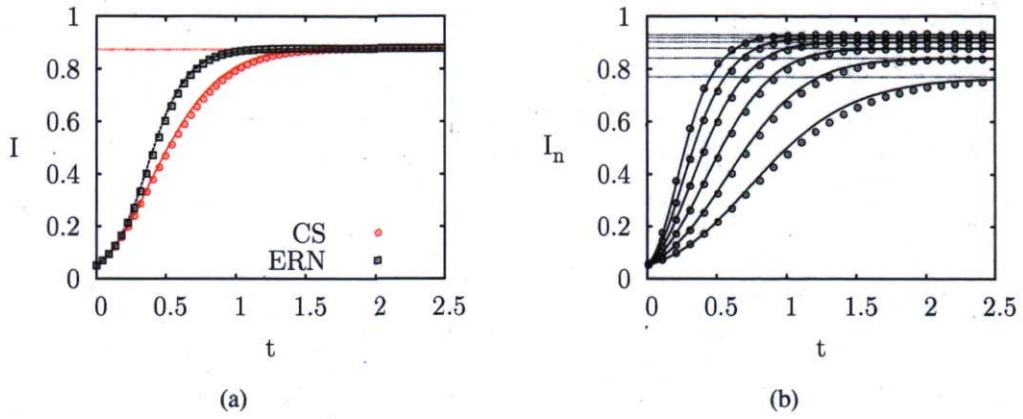



FIG. *4.5:* **Time evolution and steady state of an epidemic on community structure 2.** Comparison between analytical and numerical results on a network with general community structure defined by (4.34) for a SIS model of propagation dynamics of parameter $\lambda = \tau/r = 0.5$ under normalized time $t \to rt$. (a) time evolution of the global state (community structure in solid red and equivalent random network in dotted black) and (b) time evolution for cliques of size 10, 15, 20, 25, 30 and 35 (lowest to highest curves). All numerical results are obtained via MC simulations on over 20000 networks of 25000 nodes and are presented by their mean value. Analytical predictions for the stables states are shown in horizontal dotted lines in both figures. Note that the deviation from the predictions is bigger for the smallest cliques than for the larger ones. This is a consequence of the mean-field description which is more accurate for large systems (or, in this case, subsystems) for which standard deviations are of lesser relative importance.

### 4.7.3 Results on Newman's topology

The first topology chosen to test the formalism is the special model presented in [52], which does not allow random links and is thus obtained by setting $K_0(z) = 1$ (i.e. all links are shared within a clique). We will then use $\epsilon = 0.8$, a power-law distribution for the numbers of cliques per individual and a Poisson distribution for the numbers of individuals per clique :

$$g_m \propto m^{-1} e^{-m/1.2} \quad ; \quad p_n \propto \frac{20^n}{n!} e^{-20} . \tag{4.32}$$

This topology results in a degree distribution generated by the following function :

$$G_0(P_1(1 + (z-1)\epsilon)) = \frac{\ln(1 - e^{20(\epsilon z - \epsilon)} e^{-5/6})}{\ln(1 - e^{-5/6})} . \tag{4.33}$$

This heterogeneous distribution is shown in Fig. 4.3. To follow the propagation dynamics on an ERN, we use the first of the two options previously presented : all cliques are of size two with $\epsilon^{\text{ERN}} = 1$ and a distribution $\{g_m\}$ equivalent to (4.33).

Our results on this topology, Fig. 4.4, confirm that our formalism is indeed capable of following the time evolution of the network in structured and random topologies. Furthermore, both our numerical and analytical results support the conclusions of [34; 33; 79] as will be discussed below.

Firstly, as evident in Fig. 4.4(a), the community structure does not significantly change the stable state of the system. This conclusion is only valid when the giant components of CS and of the ERN have approximately the same size and under condition that the network is well connected. In physical terms, this means that the coupling must be sufficiently high between the subsystems, relative to the strength of the interaction (i.e. $\lambda$). If this condition is not fully met, subsets of the canonical distribution of configurations (i.e. ones with higher number of independent cliques) will have stable states under the predicted value and will decrease the mean value. The reduction of the giant component was already explained in [52]. This effect is visible in both analytical and numerical results of Fig. 4.4(a) for lower infection rate and eventually leads to a higher epidemic threshold for networks with community structure.

This particular property seems to contradict a major conclusion of [52], yet it is important to take into account that the conclusion that clustering lowers the epidemic threshold was made on networks featuring different degree distributions (see [39] for a complete discussion) and featuring degree correlation (see [48] for an analysis of correlation and clustering effects). Our results show that, given an *identical* degree distribution and zero degree correlation, the random networks will have a lower epidemic threshold than a network featuring community structure. This conclusion is intuitive because links shared in community have a higher probability of being "wasted" (i.e. of leading to another infectious node) than a random link, independently of the transmissibility. The mechanism behind this phenomenon is simple : there is a higher probability that neighbors of a new infectious individual will also be infectious if these individuals are connected in groups. This leads to a lower mean epidemic size for low infection rate and to the observed higher epidemic threshold. Note that, within the community structure effects observed here, the individual effects of clustering and degree correlation can not be separated. The demonstration given in Appendix shows that, for networks with zero degree correlation, our model always predicts a higher epidemic threshold for networks with clustering than for equivalent random networks. However, it should be emphasized that correlation effects alone have been shown to lower the percolation threshold [25]. As similar effects can take place on networks with community structure, our conclusion is not directly generalizable to networks with non-zero degree correlation.

Secondly, as seen in Fig. 4.4(b), the community structure increases the relaxation time of the system ; i.e. it slows the disease propagation towards the equilibrium. This phenomenon is also explained by the higher number of wasted links on a community structure than on the equivalent random network. These links are very frequent in social networks because

of community structure where "the friend of my friend is also my friend". When counting new possible infections on networks with exactly the same degree distribution, the number of second neighbors will be higher in a random network than on a community structure, because the neighbors of my neighbor may have already been counted as my neighbor in the CS network. This results in a slower propagation and a typically higher epidemic threshold.

Finally, note that the shift observed in the epidemic threshold is not always as small as seen on Fig. 4.4(a). For example, a topology with $G'(1) \simeq 0.365$ and $\nu = 5$ yields $\lambda_c^{\mathrm{CS}} = 5/4 \cdot \lambda_c^{\mathrm{ERN}}$. This particular case was verified by MC simulations.

### 4.7.4 Results on a general topology

As a second test to our formalism, we use $\epsilon = 0.8$ and the following distributions :

$$g_m \propto \frac{e^{-4m}}{m}; \quad p_n \propto \frac{20^n}{n!}e^{-20}; \quad k_l \propto \frac{e^{-l}}{l} \tag{4.34}$$

which result in the second degree distribution shown in Fig. 4.3 and generated by :

$$\frac{\ln(1 - e^{20(\epsilon z - \epsilon)}e^{-5/6})}{\ln(1 - e^{-5/6})} \frac{\ln(1 - ze^{-1})}{\ln(1 - e^{-1})} . \tag{4.35}$$

In this case, the ERN are obtained by using cliques of size one and fitting the degree distribution with the random links generated by $K_0(z)$. The results obtained on this second topology are presented in Fig. 4.5. They not only confirm the quality of our treatment, but also earlier conclusions. The propagation slow-down is stronger in the time evolution featured in Fig. 4.5(a) than in the case observed in Fig. 4.4(b), because the topology used produces a much higher proportion of intra-clique links for a given individual, and consequently, a higher fraction of wasted links. It is believed that this effect could be studied using percolation theory with a quantification of CS, such as the modularity concept introduced by Newman and Girvan in [57].

## 4.8 Conclusion

What may well be the single most important contribution of this paper is the philosophy upon which the formalism is based. An effective dynamical description of complex networks can be obtained by a mean-field approach using a compartmentalisation of both the networks' elements (e.g. individuals or nodes) and of their recurrent topological patterns (e.g. cliques or substructures) in classes of homogeneous state and behaviour. It has been shown that a particular topology, the community structure, can be solved with this method. Furthermore, the approach can also describe random topology in the limit of the most elementary patterns

possible. Hence, it is reasonable to assert that other complex topologies may be treated in a similar manner.

More precisely, our analytical results confirm previous numerical simulations on the effects of community structure in propagation dynamics : in comparison to equivalent random networks, the structured systems feature longer relaxation times (i.e. slower propagation) and generally higher epidemic thresholds.

An especially interesting avenue to explore would be to direct the formalism towards more epidemiologically oriented applications with a generalization to other propagation model (see for example [7]). Furthermore, in an epidemic context, taking the topology of social network into account allows precise emulation of real intervention scenarios which are often based on groups of individuals (e.g. school closings and vaccination of public health workers both correspond to interventions on given cliques).

Other applications of our formalism are possible in various models of dynamics and topologies. Of particular interest is the application of our formalism to dynamical networks (e.g. [63; 44]). This may help in gaining insights on the emergence and the stability of social structure.

## 4.9 Appendix A : Community structure, without degree correlation, raises the epidemic threshold

This paper has shown that our model can describe propagation phenomena on network with community structure as well as network with random topology. Using the analytic solution for the epidemic threshold on Newman's topology, it is possible to show that, given two networks with identical degree distributions and zero degree correlation, but where one is completely random while the other features community structure (and therefore clustering), the latter will have a higher epidemic threshold.

First of all, degree correlation refers to situations where, given a random link in the network, the knowledge of the excess degree of one of its nodes influences the probability distribution for the excess degree of the other. For Newman's model, it was shown in [58] that the probability $e_{jk}$ that a given link joins two nodes of excess degree $j$ and $k$ can be calculated as follows. We first write :

$$e_{jk} = \frac{1}{N} \sum_n p_n n(n-1) P(j,k|n) \,, \tag{4.36}$$

where $n(n-1)$ is the number of potential degrees in a clique of size $n$, $N$ is a normalization factor corresponding to the total number of potential links in the network and $P(j,k|n)$ is the

probability that a link within a clique of size $n$ joins two nodes of excess degree $j$ and $k$. This probability can be calculated by separating $j$ in $j_{\text{in}}$ and $j_{\text{out}}$, respectively the excess links shared within and outside of the considered clique, and doing the same for $k$. We can now write :

$$P(j,k|n) = \sum_{j_{\text{in}}} \binom{n-2}{j_{\text{in}}} \epsilon^{j_{\text{in}}} (1-\epsilon)^{n-2-j_{\text{in}}} P(j_{\text{out}}) + \sum_{k_{\text{in}}} \binom{n-2}{k_{\text{in}}} \epsilon^{k_{\text{in}}} (1-\epsilon)^{n-2-k_{\text{in}}} P(k_{\text{out}}) \,, \quad (4.37)$$

where $P(j_{\text{out}})$ and $P(k_{\text{out}})$ are the probabilities that the nodes have $j_{\text{out}}$ and $k_{\text{out}}$ links outside the clique of size $n$. These two probabilities are simply generated by the PGFs composition $G_1(P_1(1+(z-1)\epsilon))$. Now, because both $k$ and $j$ must be calculated with one clique in common where they both have $n-2$ potential excess neighbors, we can write the set of $\{e_{jk}\}$ in terms of the following PGF :

$$\begin{aligned} \sum_{jk} e_{jk} x^j y^k =& P_2\left((1+(x-1)\epsilon)(1+(y-1)\epsilon)\right) \\ & \times G_1\left(P_1(1+(x-1)\epsilon)\right) G_1\left(P_1(1+(y-1)\epsilon)\right) \,, \end{aligned} \quad (4.38)$$

where $P_2(z) \equiv \left[P_0''(1)\right]^{-1} \sum_n n(n-1) z^{n-2}$. For a random network, it is easily obtained that $e_{jk}$ is simply the product of the two independent probabilities of having nodes of excess degree $j$ and $k$. Thus, by differentiating the degree distribution PGF (4.30) to obtain the excess degree distribution, we find :

$$\begin{aligned} \sum_{jk} e_{jk}^{ERN} x^j y^k =& P_2(1+(x-1)\epsilon)\, G_1\left(P_1(1+(x-1)\epsilon)\right) \\ & \times P_2(1+(y-1)\epsilon)\, G_1\left(P_1(1+(y-1)\epsilon)\right) \,. \end{aligned} \quad (4.39)$$

For expressions (4.38) and (4.39) to be equivalent, the following condition must be satisfied :

$$P_2\left((1+(x-1)\epsilon)(1+(y-1)\epsilon)\right) = P_2(1+(x-1)\epsilon)\, P_2(1+(y-1)\epsilon) \,. \quad (4.40)$$

We want to compare two networks sharing exactly the same degree distribution and degree correlation. Equation (4.40) gives us the condition for which two networks with identical degree distributions, one featuring community structure and the other random topology, will have the same degree correlation. It is easy to conclude that the distribution of individuals per clique, in order to respect Eq. (4.40), can only be given by :

$$p_n = \delta_{n,\nu} \,, \quad (4.41)$$

where $\nu$ is an arbitrary positive integer. In other words, all structures must be the same size. This limitation comes from the way we construct our random networks. Because by simply matching degrees generated from a given distribution, the knowledge of one neighbor's degree does not give any information concerning the other neighbor's degree. Note that $G_0(z)$

and $\epsilon$ are totally free, so that the heterogeneity of the degree distribution is not entirely compromised.

We will now compare two networks with zero degree correlations. The first is random with $p_n^{\mathrm{ERN}} = \delta_{n,2}$ and $\epsilon^{\mathrm{ERN}} = 1$ while the other exhibits community structure with $p_n^{\mathrm{CS}} = \delta_{n,\nu}$ with $\nu > 2$ and $\epsilon^{\mathrm{CS}} \equiv \epsilon \in [0, 1]$. The two networks have exactly the same degree distribution, which means that $G_0^{\mathrm{ERN}}(z) = G_0^{\mathrm{CS}}\left(P^{\mathrm{CS}}(1 + (z-1)\epsilon)\right)$. Using Eq. (4.28), we can easily write the epidemic threshold for the random network :

$$\lambda_c^{\mathrm{ERN}} = \frac{1}{\frac{d}{dz}G_1^{\mathrm{ERN}}(1)} \equiv \frac{1}{\epsilon\left[(\nu-2) + \mu_1(\nu-1)\right]}, \tag{4.42}$$

where the last expression uses the PGFs of the structured network in which $\mu_1 = \frac{d}{dz}G_1^{\mathrm{CS}}(1)$ is the mean number of excess cliques per individual. We will now insert expression (4.42) in the epidemic threshold condition (4.26) of the network with community structure. Because all terms in the polynomial are positive, we expect to find an expression greater than unity if (4.42) is higher than the threshold for CS, equal to one if the threshold remains the same or lesser than unity if the threshold for the ERN is actually lower than that for CS. To prove the latter case, for arbitrary $\nu$, $\epsilon$ and $\{g_m\}$, we simply demonstrate the following inequality written from (4.26) using (4.42) :

$$\sum_{i=1}^{\nu-1} \frac{\mu_1(\nu-1)!}{(\nu-i-1)!}\left[(\nu-2) + \mu_1(\nu-1)\right]^{-i} < 1 . \tag{4.43}$$

Further, it can be shown that the derivative of (4.43) in $\mu_1$ is always positive. This provides us with an upper bound for (4.43) in the limit $\mu_1 \to \infty$. Using l'Hôpital's rule, we thus find :

$$\lim_{\mu_1\to\infty} \sum_{i=1}^{\nu-1} \frac{\mu_1(\nu-1)!}{(\nu-i-1)!}\left[(\nu-2) + \mu_1(\nu-1)\right]^{-i} = 1 . \tag{4.44}$$

This indicates that the two networks with zero degree correlation, one featuring community structure and one an equivalent random network, will have the same threshold in the limit of infinite mean number of excess cliques per individual or if $\nu = 2$. Otherwise, because the derivative of the polynomial in $\mu_1$ was shown to be positive, finite $\mu_1$ and $\nu > 2$ imply a higher threshold for the structured network.

# Chapitre 5

# Conclusion et perspectives : un nouveau paradigme

Pour paraphraser de nouveau Herbert Simon, mentionnons que le but du présent mémoire n'était pas de supposer une connexion entre le choix de mots d'un auteur, la croissance de l'Internet et les voies de transmission d'une épidémie de grippe, mais simplement de mettre en lumière le fait que le même processus stochastique procure un modèle satisfaisant de ces différents phénomènes. Comme avait écrit par la suite Benoît Mandelbrot dans sa première réponse à l'article de Simon sur l'attachement préférentiel [43], ceci *s'insère dans l'un des buts universels de l'explication scientifique, qui est de réduire la complexité à la simplicité ; même si cette simplicité est à un niveau difficile à comprendre.*

Dans les cas qui nous intéressent, les réseaux complexes et les systèmes libres d'échelle en tout genre, la physique statistique offrait des outils bien adaptés à notre objectif, car ces systèmes sont composés d'un grand nombre d'éléments organisés de façon ni complètement ordonnée, ni tout à fait aléatoire. En décrivant leur dynamique[1] à l'aide de processus stochastiques, on est arrivé à mieux les comprendre (e.g. une possible origine pour l'auto-similarité des réseaux complexes, voir Chap. 2), à prédire certaines de leurs propriétés (e.g. la dynamique de peloton, voir Chap. 3) ainsi que le comportement des dynamiques qu'ils peuvent soutenir (e.g. propagation d'épidémies sur réseaux sociaux, voir Chap. 4). Au-delà de l'observation, de la compréhension, de la modélisation et de la prédiction, la dernière étape de la maîtrise d'un système complexe est de pouvoir le contrôler.

Dans cette optique, on peut vouloir tester, par exemple, l'effet d'interventions sur un réseau social en cas d'épidémies. Ce genre d'études pourrait éventuellement apporter de

[1] Soit la dynamique *du* système, soit une dynamique *sur* le système.

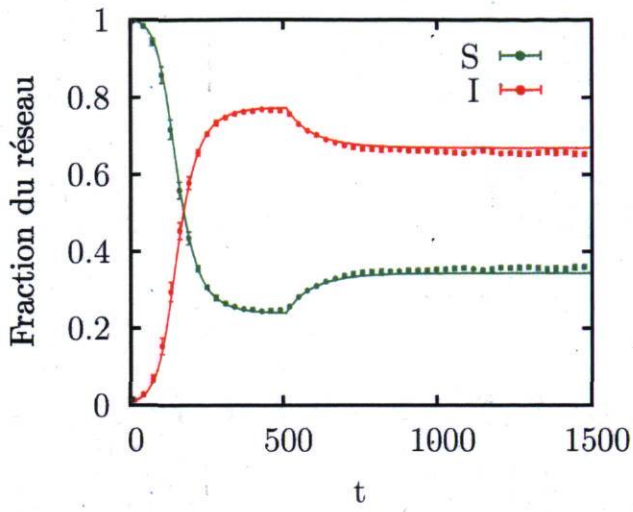



FIG. *5.1:* **Simulation de quarantaine sur un réseau social.** Évolution temporelle de l'état d'un réseau à distributions binomiales de moyennes 3 (cliques par individu) et 20 (individus par clique) dans une dynamique SIS avec $\tau = 0.001$, $r = 0.005$ et $\epsilon = 0.5$ telle qu'obtenue par simulation et par l'intégration du formalisme présenté au Chap. 4. Les cliques de taille supérieure à 23 sont *fermées* à $t = 500$. Ce simple cas émule la fermeture de certains groupes ou institutions en cas de pandémie. Les termes représentant les contacts à l'intérieur de certaines cliques sont ainsi annulés et les termes de couplages $C_{n>c}$ correspondant sont proportionnellement réduits. L'utilité est de pouvoir, par exemple, prédire l'efficacité de la fermeture d'écoles pour ralentir une épidémie de grippe, en autant d'avoir accès à des données statistiques précises sur la structure communautaire du système réel.

précieuses informations pour guider la prise de décision au niveau de la santé publique. Notre modèle peut ainsi être utilisé pour simuler des quarantaines en temps de pandémie (voir Fig. 5.1). Cet exercice illustre bien à quel point la recherche théorique et la recherche appliquée évoluent main dans la main en science des réseaux.

Cela étant dit, pour l'instant nos résultats ont un attrait beaucoup plus théorique que pratique. En fait, les travaux exposés dans ce mémoire simplifient et élargissent notre compréhension des réseaux complexes en unifiant leur description sous le concept de la *structure communautaire*. Il a été démontré comment ce nouveau paradigme simplifiait la description de l'organisation des réseaux complexes. Notre portée n'étant jamais plus large que la profondeur de ce que l'on comprend, ce nouveau paradigme pourrait s'avérer être la pierre d'assise de nouveaux efforts pour mieux décrire et utiliser les réseaux complexes.

À ce sujet, on a mentionné en introduction que l'épidémiologie sur réseaux jouait un double rôle dans ce mémoire. D'une part, il s'agissait d'une belle application du nouveau paradigme d'organisation sur un problème de haute importance. Et d'autre part, il s'agissait

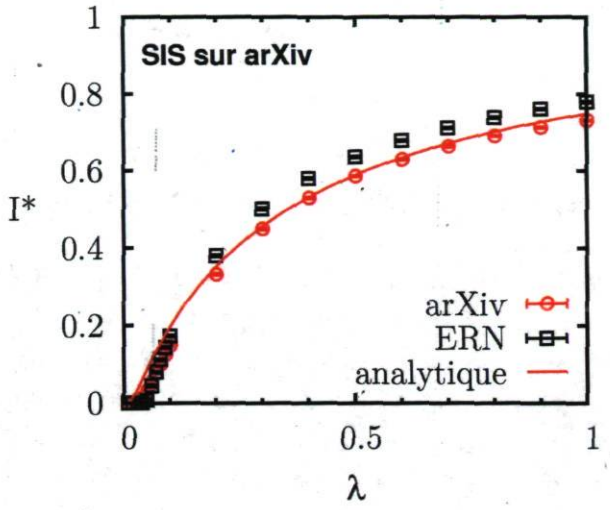



FIG. *5.2:* **Prédiction d'une épidémie sur *arXiv*.** Courbe analytique : solution de l'équation (4.24), utilisant les fonctions génératrices (3.11) et (3.12) avec les paramètres de SPA utilisés à la Fig. 2.3(a). Cette prédiction est comparée aux résultats des simulations d'un modèle SIS de propagation d'épidémie sur le réseau du *cond-mat arXiv*. De plus, les résultats obtenus sur des réseaux aléatoires équivalents (produits à partir du modèle présenté à la Sec. 1.2.3) sont utilisés pour illustrer l'effet de la structure du réseau réel. Les courbes correspondent à la fraction infectée du réseau ($I^*$) lorsque la maladie atteint son équilibre entre infections et rétablissements en fonction du ratio du taux d'infection $\tau$ et du taux de rétablissement $r$ ($\lambda = \tau/r$).

en quelque sorte d'une ultime validation de notre principe d'organisation basé sur la structure communautaire. En effet, si l'on réussit à bien décrire comment une épidémie se propage sur un réseau réel, on peut avoir bon espoir d'avoir réussi à décrire les propriétés topologiques importantes de ce réseau (ou au moins celles importantes par rapport à cette dynamique).

Rappelons maintenant la petite expérience effectuée en toute fin du Chap. 1 où les simulations d'épidémie (ou de la propagation d'une information) sur l'archive de matière condensée *cond-mat arXiv* ont été comparées à des simulations sur un réseau aléatoire équivalent. On avait alors pu illustrer l'effet de l'organisation de l'*arXiv*, i.e. de la topologie du réseau, en comparant les deux ensembles de résultats. Maintenant que nous avons obtenu un modèle de cette organisation, fait une description analytique de la structure résultante et de la dynamique qui s'y propage, on peut comparer les résultats de nos efforts avec les simulations initiales. Ceci est présenté à la Fig. 5.2.

Mentionnons d'abord que le fait de considérer une dynamique SIS plutôt qu'un problème de percolation favorise ce type de description par champs moyens (évolution moyenne d'un état moyen sur la structure moyenne) comme toute simulation finie par visiter un grand en-

semble d'états possibles. Cela étant dit, ces résultats, qui regroupent les travaux des trois principaux chapitres du mémoire, confirment tout de même la qualité de nos modèles. L'aspect le plus satisfaisant de ce constat est que cela indique que les méthodes développées sont adéquates pour la science des réseaux. On peut ensuite espérer pouvoir obtenir de meilleures solutions, soit dans le cadre d'une dynamique de propagation ou d'un autre problème, en perfectionnant notre description de l'organisation des réseaux réels.

Par exemple, la légère différence entre les simulations sur l'*arXiv* et la prédiction de nos modèles pour de grand $\lambda$ est potentiellement simplement due aux corrélations observées au Chap. 3, mais reflète peut être aussi l'existence d'autres niveaux d'organisation. C'est dans ce contexte plus ambitieux que s'inscrira la suite de ce projet de recherche. On visera à obtenir une meilleure description des systèmes complexes en unifiant des descriptions simplifiées de tous les niveaux d'organisation dont ils peuvent être composés.

Le but ultime de cette étude est de pouvoir catégoriser les systèmes complexes en regroupant leurs propriétés en diverses classes d'universalité. Il s'agit en somme d'utiliser les outils de la physique statistique pour appliquer aux réseaux complexes de nature sociale, écologique, biologique ou technologique, des méthodes d'analyse ayant fait leurs preuves en thermodynamique, physique de l'état solide et matière condensée. Voilà un des grands objectifs de la *science des réseaux*, et ce mémoire est, je l'espère, un pas dans la bonne direction.

# Bibliographie


[1] Y.-Y. Ahn, J. P. Bagrow et S. Lehmann, *Link communities reveal multiscale complexity in networks*, Nature, 466 (2010), p 761.

[2] R. Albert et A.-L. Barabási, *Topology of evolving networks : Local events and universality*, Phys. Rev. Lett., 85 (2000), p 5234.

[3] R. Albert, H. Jeong et A.-L. Barabási, *Error and attack tolerance of complex networks*, Nature, 406 (2000), p 378.

[4] A. Allard, P.-A. Noël, L. J. Dubé et B. Pourbohloul, *Heterogeneous bond percolation on multitype networks with an application to epidemic dynamics*, Phys. Rev. E, 79 (2009), p 036113.

[5] R. M. Anderson et R. M. May, *Infectious Disease of Humans : Dynamics and Control*, Oxford University Press, 1991.

[6] F. Ball, *Stochastic and deterministic models for SIS epidemics among a population partitioned into households*, Math. Biosci., 156 (1999), p. 41–68.

[7] F. Ball, D. Sirl et P. Trapman, *Analysis of a stochastic SIR epidemic on a random network incorporating household structure*, Math. Biosci., 224(2) (2010), p. 53–73.

[8] A.-L. Barabási, *Scale-free networks : A decade and beyond*, Science, 325 (2009), p 412.

[9] A.-L. Barabási et R. Albert, *Emergence of scaling in random networks*, Science, 286 (1999), p 509.

[10] A. Barrat, M. Barthélemy et A. Vespignani, *Dynamical Processes on Complex Networks*, Cambridge University Press, 2008.

[11] S. Bernhardsson, L. E. C. da Rocha et P. Minnhagen, *The meta book and size-dependent properties of written language*, New Journal of Physics, 11 (2009), p 123015.

[12] K. Christensen et N. R. Moloney, *Complexity and Criticality*, Imperial College Press, 2005.

[13] A. Clauset, C. Moore et M. E. J. Newman, *Hierarchical structure and the prediction of missing links in networks*, Nature, 453 (2008), p 98.

[14] A. Clauset, M. E. J. Newman et C. Moore, *Finding community structure in very large networks*, Phys. Rev. E, 70 (2004), p 066111.

[15] S. N. Dorogovtsev et J. F. F. Mendes, *Evolution of Networks : From Biological Nets to the Internet and WWW*, Oxford University Press, 2003.

[16] J. Doyle et J. Carlson, *Power laws, highly optimized tolerance, and generalized source coding*, Phys. Rev. Lett., 84 (2000), p 5656.

[17] P. Erdös et A. Rényi, *On random graphs. I*, Publicationes Mathematicae, 6 (1959), p 290–297.

[18] L. Euler, *Solutio problematis ad geometriam situs pertinentis*, Commentarii academiae scientiarum Petropolitanae, 8 (1741), p. 128–140.

[19] M. J. Ferrari, S. Bansal, L. A. Meyers et O. N. Bjørnstad, *Network frailty and the geometry of herd immunity*, Proc. R. Soc. B, 273 (2006), p 2743–2748.

[20] C. Furusawa et K. Kaneko, *Zipf 's law in gene expression*, Phys. Rev. Lett., 90 (2003), p 088102.

[21] G. Ghoshal, L. Sander et I. Sokolov, *SIS epidemics with household structure : the self-consistent field method*, Math. Biosci., 190 (2004), p 71–85.

[22] R. Gibrat, *Les inégalités économiques*, Librairie du Recueil Sirey, 1931.

[23] M. Girvan et M. E. J. Newman, *Community structure in social and biological networks*, Proc. Natl. Acad. Sci., 99 (2002), p 7821.

[24] J. P. Gleeson, *Bond percolation on a class of clustered random networks*, Phys. Rev. E, 80 (2009), p 036107.

[25] J. P. Gleeson, S. Melnik et A. Hackett, *How clustering affects the bond percolation threshold in complex networks*, Phys. Rev. E, 81 (2010), p 066114.

[26] C. Godrèche et J. M. Luck, *On leaders and condensates in a growing networks*, J. Stat. Mech., (2010), p P07031.

[27] 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 (2003), p 065103.

[28] H. S. Heaps, *Information Retrieval : Computational and Theoretical Aspects*, Academic Press, 1978.

[29] L. Hébert-Dufresne, A. Allard, V. Marceau, P.-A. Noël et L. J. Dubé, *Structural preferential attachment : Network organization beyond the link*, Phys. Rev. Lett., 107 (2011), p 158702.

[30] 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 (2010), p 036115.

[31] T. Heimo, J. Saramäki, J.-P. Onnela et K. Kaski, *Spectral and network methods in the analysis of correlation matrices of stock returns*, Physica A, 383 (2007), p 147–151.

[32] A. Hertz, *L'Agrapheur*, Presses Internationales Polytechnique, 2010.

[33] D. Hiebeler, *Moment equations and dynamics of a household sis epidemiological model*, Bull. Math. Biol., 68 (2006), p. 1345–1333.

[34] W. Huang et C. Li, *Epidemic spreading in scale-free networks with community structure*, J. Stat. Mech., (2007), p P01014.

[35] B. Karrer, E. Levina et M. E. J. Newman, *Robustness of community structure in networks*, Phys. Rev. E, 77 (2008), p 046119.

[36] B. Karrer et M. E. J. Newman, *Random graphs containing arbitrary distributions of subgraphs*, Phys. Rev. E, 82 (2010), p 066118.

[37] M. Keeling, *The implications of network structure for epidemic dynamics*, Theor. Popul. Biol., 67 (2005), p. 1–8.

[38] M. J. Keeling et K. T. D. Eames, *Networks and epidemic models*, J. R. Soc. Interface, 2 (2005), p 295–307.

[39] I. Z. Kiss et D. M. Green, *Comment on "Properties of highly clustered networks"*, Phys. Rev. E, 78 (2008), p 048101.

[40] D. E. Knuth, *The Stanford GraphBase : A Platform for Combinatorial Computing*, Addison-Wesley Professional, 1993.

[41] P. Krapivsky et S. Redner, *Statistics of changes in lead node in connectivity-driven networks*, Phys. Rev. Lett., 89 (2002), p 258703.

[42] L. Lovász, *Combinatorics, Complexity, and Chance, A Tribute to Dominic Welsh*, Oxford Univ. Press, 2007, chap. Connection matrices, p. 179–190.

[43] B. Mandelbrot, *A note on a class of skew distribution functions : Analysis and critique of a paper by H.A. Simon*, Information and Control, 2 (1959), p 90.

[44] 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 (2010), p 036116.

[45] ——, *Modeling the dynamical interaction between epidemics on overlay networks*, Phys. Rev. E, 84 (2011), p 026105.

[46] M. MARDER, *Dynamics of epidemics on random networks*, Phys. Rev. E, 75 (2007), p 066103.

[47] R. M. MAY, *Network structure and the biology of populations*, Trends Ecol. Evol., 21 (2006), p. 394–399.

[48] J. C. MILLER, *Percolation and epidemics in random clustered network*, Phys. Rev. E, 80 (2009), p 020901.

[49] B. R. MORIN, *Explicit solutions to the continuous time Albert-Barabási scale-free model*, arXiv, (2011), p 1105.0882.

[50] M. E. J. NEWMAN, *Random graphs with arbitrary degree distributions and their applications*, Phys. Rev. E, 64 (2001), p 026118.

[51] ——, *Spread of epidemic disease on networks*, Phys. Rev. E, 66 (2002), p 016128.

[52] ——, *Properties of highly clustered networks*, Phys. Rev. E, 68 (2003), p 026121.

[53] ——, *Detecting community structure in networks*, Eur. Phys. J. B, 38 (2004), p 321–330.

[54] ——, *Power laws, Pareto distributions and Zipf's law*, Contemporary Physics, 46 (2005), p 323.

[55] ——, *Modularity and community structure in networks*, PNAS, 103 (2006), p. 8577 – 8582.

[56] ——, *Random graphs with clustering*, Phys. Rev. Lett., 103 (2009), p 058701.

[57] M. E. J. NEWMAN ET M. GIRVAN, *Finding and evaluating community structure in networks*, Phys. Rev. E, 69 (2004), p 026113.

[58] M. E. J. NEWMAN ET J. PARK, *Why social networks are different from other types of networks*, Phys. Rev. E, 68 (2003), p 036122.

[59] 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 (2009), p 026101.

[60] J.-P. ONNELA, A. CHAKRABORTI, K. KASKI, J. KERTÉSZ ET A. KANTO, *Dynamics of market correlations : Taxonomy and portfolio analysis*, Phys. Rev. E, 68 (2003), p 056110.

[61] G. PALLA, A.-L. BARABÁSI ET T. VICSEK, *Quantifying social group evolution*, Nature, 446 (2007), p 664.

[62] G. Palla, I. Derényi, I. Farkas et T. Vicsek, *Uncovering the overlapping community structure of complex networks in nature and society*, Nature, 435 (2005), p 814.

[63] G. Palla, P. Pollner, A.-L. Barabási et T. Vicsek, *Adaptive Networks*, Springer, 2009, chap. 2, p. 11–50.

[64] R. Parshani, S. Carmi et S. Havlin, *Epidemic threshold for the susceptible-infectious-susceptible model on random networks*, Phys. Rev. Lett., 104 (2010), p 258701.

[65] R. Pastor-Satorras et A. Vespignani, *Epidemic spreading in scale-free networks*, Phys. Rev. Lett., 86 (2001), p. 3200–3203.

[66] M. Pautasso et M. J. Jeger, *Epidemic threshold and network structure : The interplay of probability of transmission and of persistence in small-size directed networks*, Ecol. Compl. 5, 5 (2008), p. 1–8.

[67] V. Plerou, L. A. Amaral, P. Gopikrishnan, M. Meyer et H. E. Stanley, *Similarities between the growth dynamics of university research and of competitive economic activities*, Nature, 400 (1999), p 433.

[68] J. M. Pujol, J. Béjar et J. Delgado, *Clustering algorithm for determining community structure in large networks*, Phys. Rev. E, 74 (2006), p 016107.

[69] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai et A.-L. Barabási, *Hierarchical organization of modularity in metabolic networks*, Science, 297 (2002), p 1551.

[70] D. Rybski, S. V. Buldyrev, S. Havlin, F. Liljeros et H. A. Makse, *Scaling laws of human interaction activity*, Proc. Natl. Acad. Sci., 106 (2009), p 12640.

[71] M. D. Shirley et S. P. Rushton, *The impacts of network topology on disease spread*, Ecol. Compl., 2 (2005), p. 287–299.

[72] H. A. Simon, *On a class of skew distribution functions*, Biometrika, 42 (1955), p 425.

[73] ——, *Models of Man*, John Wiley & Sons, 1961.

[74] C. Song, S. Havlin et H. Makse, *Self-similarity of complex networks*, Nature, 433 (2005), p 392.

[75] ——, *Origins of fractality in the growth of complex networks*, Nature Physics, 2 (2006), p 275.

[76] V. Spirin et L. A. Mirny, *Protein complexes and functional modules in molecular networks*, PNAS, 100 (2003), p 12123–12128.

[77] J. Travers et S. Milgram, *An experimental study of the small world problem*, Sociometry, 32 (1969), p 425.

[78] E. Volz, *SIR dynamics in random networks with heterogeneous connectivity*, J. Math. Biol., 56 (2008), p. 293–310.

[79] D. J. Watts, R. Muhamad, D. C. Medina et P. S. Dodds, *Multiscale, resurgent epidemics in a hierarchical metapopulation model*, PNAS, 102 (2005), p 11157–11162.

[80] D. J. Watts et S. H. Strogatz, *Collective dynamics of 'small-world' networks*, Nature, 393 (1998), p 440.

[81] H. S. Wilf, *generatingfunctionology*, Academic Press, Inc., 1990.

[82] J. Willis et G. Yule, *Some statistics of evolution and geographical distribution in plants and animals, and their significance*, Nature, 109 (1922), p 177.

[83] G. U. Yule, *A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F.R.S.*, Phil. Trans. R. Soc. Lond. B, 213 (1925), p 21.

[84] G. K. Zipf, *Human Behavior and the Principle of Least Effort*, Addison-Wesley Press, 1949.