

# **Processus de contagion sur réseaux complexes au-delà des interactions dyadiques**

**Thèse**

**Guillaume St-Onge**

Sous la direction de:

Antoine Allard, directeur de recherche  
Laurent Hébert-Dufresne, codirecteur de recherche

# Résumé

Alors que la pandémie de COVID-19 affecte le monde depuis presque deux ans, il va sans dire qu'une meilleure compréhension des processus de contagion, de leur évolution et des effets des mesures de contrôle est essentielle pour réduire leur impact sur la société. Le cadre théorique pour la modélisation des processus de contagion est très général et permet, bien entendu, de décrire la propagation des maladies infectieuses causées par des agents pathogènes (virus, bactéries, parasites, etc.), mais aussi la propagation des rumeurs et de la désinformation. Peu importe la nature du processus, la transmission s'effectue de proche en proche grâce aux interactions entre les individus. Par conséquent, la structure sociale *complexe* des populations, qui n'est ni parfaitement ordonnée, ni complètement aléatoire, joue un rôle de premier plan.

Dans cette thèse, nous étudions les processus de contagion sur réseaux, où les individus et les interactions entre ces individus sont représentés par des noeuds et des liens respectivement. Nous utilisons une approche théorique principalement basée sur la physique statistique et la dynamique non linéaire. Nous nous concentrons plus spécifiquement sur les *réseaux d'ordre supérieur*, lesquels mettent les interactions de *groupe* à l'avant-plan. Notre analyse va donc au-delà des interactions dyadiques. Bien plus qu'une reformulation mathématique de la structure, cette perspective est primordiale pour obtenir une compréhension plus complète de la phénoménologie des processus de contagion.

Nous démontrons l'importance des interactions de groupe à l'aide de trois résultats principaux. D'abord, nous caractérisons un phénomène de *localisation mésoscopique* : pour certaines structures hétérogènes, la propagation persiste uniquement dans les groupes de grande taille. Ce phénomène a notamment une incidence sur l'effet des mesures de contrôle visant à prohiber les regroupements au-delà d'une certaine taille, à l'instar de ce qui fut instauré pour endiguer la pandémie de COVID-19. Ensuite, nous étudions un modèle où les individus doivent accumuler une dose infectieuse minimale pour devenir infectés. Nous montrons qu'une structure d'ordre supérieur et des temps d'exposition hétérogènes induisent une probabilité d'infection non linéaire *universelle*. L'épidémie résultante peut alors croître de manière *super-exponentielle* en fonction du temps. Finalement, nous poussons plus en profondeur l'analyse des processus de contagion non linéaire. Dans ce contexte, nous montrons que les groupes peuvent avoir plus d'importance que les individus ultra-connectés pour qu'une épidémie ou un phénomène social envahissent le plus rapidement possible une population.

# Abstract

After almost two years into the COVID-19 pandemic, it is clear that a better understanding of contagion processes, their evolution, and the impact of control measures is essential to reduce their burden on society. The theoretical framework for the modeling of contagion is quite general. It can describe the spread of pathogens causing diseases (viruses, bacteria, parasites, etc.), but also the spread of rumors and disinformation. Irrespective of the nature of the underlying process, the contagion evolves through local interactions between the individuals. Consequently, the *complex* social structure of populations, which is neither perfectly ordered nor completely random, plays a fundamental role in shaping spreading.

In this thesis, we study contagion processes on networks where individuals and the interaction between them are represented by nodes and edges respectively. We use a theoretical approach based on statistical physics and nonlinear dynamics. We focus on *higher-order networks*, putting group interactions—beyond pairwise interactions—at the forefront. More than a mere mathematical generalization, we find this perspective is paramount to obtain a complete picture of the phenomenology of contagion dynamics.

We demonstrate the importance of group interactions through three principal results. First, we characterize a *mesoscopic localization* phenomenon where the contagion thrives only in large groups for certain types of heterogeneous structure. This phenomenon significantly affects the results of interventions like the cancelation of events larger than a critical size, similar to the measures being used to limit the spreading of COVID-19. Second, we study a model where individuals must accumulate a minimal infective dose to become infected. We show that a higher-order structure and heterogeneous exposure induce a *universal* nonlinear infection probability. The epidemic size can then grow super-exponentially with time. Finally, with a more in-depth analysis of nonlinear contagions, we show that groups can be more influential than hubs (super-connected individuals) to maximize the early spread of an epidemic.

# Table des matières

Résumé	ii
Abstract	iii
Table des matières	iv
Liste des tableaux	vii
Liste des figures	viii
Liste des abréviations	x
Liste des contributions	xi
Notation	xiii
Remerciements	xvi
Avant-propos	xviii
Introduction	1
<b>I Notions de base pour les processus de contagions sur réseaux</b>	<b>5</b>
<b>1 Représentation en réseau des systèmes complexes</b>	<b>6</b>
1.1 Graphe : interactions dyadiques . . . . .	7
1.2 Réseaux d'ordre supérieur . . . . .	10
1.3 Réseaux aléatoires . . . . .	14
1.4 Reconstruction et propriétés statistiques des systèmes sociaux . . . . .	20
<b>2 Processus de contagion</b>	<b>26</b>
2.1 Cadre théorique et formalismes compartimentaux . . . . .	26
2.2 Contagions simples . . . . .	28
2.3 Contagions complexes . . . . .	31
2.4 Formalismes compartimentaux pour les contagions sur réseaux . . . . .	36
2.5 Transition de phase étendue et localisation . . . . .	42
<b>II Contagions simples</b>	<b>45</b>
<b>3 Localisation mésoscopique des processus de contagion</b>	<b>46</b>
3.1 Avant-propos . . . . .	47
3.2 Résumé . . . . .	48
3.3 Abstract . . . . .	49

3.4	Introduction . . . . .	49
3.5	Group-based SIS model . . . . .	51
3.6	Mesoscopic localization . . . . .	57
3.7	Discussion . . . . .	63
3.8	Appendix : Detailed development for the epidemic threshold . . . . .	64
3.9	Appendix : Bounds on the epidemic threshold . . . . .	65
3.10	Appendix : Generation of networks with cliques . . . . .	66
3.11	Appendix : Validation with Monte Carlo simulations . . . . .	67
3.12	Appendix : Scaling the transmission rate with group size . . . . .	68
3.13	Appendix : Localized portion of the bifurcation diagram . . . . .	70
<b>4</b>	<b>Localisation et mesures de contrôle</b>	<b>72</b>
4.1	Avant-propos . . . . .	73
4.2	Résumé . . . . .	73
4.3	Abstract . . . . .	73
4.4	Introduction . . . . .	73
4.5	Group-based framework . . . . .	74
4.6	Results . . . . .	76
4.7	Conclusion and perspectives . . . . .	79
<b>III</b>	<b>Contagions complexes</b>	<b>81</b>
<b>5</b>	<b>Universalité des contagions non linéaires en loi de puissance</b>	<b>82</b>
5.1	Avant-propos . . . . .	83
5.2	Résumé . . . . .	84
5.3	Abstract . . . . .	85
5.4	Introduction . . . . .	85
5.5	Universal infection kernel from heterogeneous exposure . . . . .	87
5.6	Epidemic spreading with nonlinear infection kernel . . . . .	88
5.7	Conclusion . . . . .	90
5.8	Appendix : Nonlinear infection kernel with general dose distribution . . . . .	91
5.9	Appendix : Characterization of the phase transition . . . . .	100
5.10	Appendix : Tracking dynamical correlations for quenched systems . . . . .	103
5.11	Appendix : Comparison with simulations . . . . .	104
5.12	Appendix : Code and data . . . . .	107
<b>6</b>	<b>Sur l'importance des groupes influents pour les contagions non linéaires</b>	<b>108</b>
6.1	Avant-propos . . . . .	109
6.2	Résumé . . . . .	110
6.3	Abstract . . . . .	111
6.4	Introduction . . . . .	111
6.5	Hypergraph contagion model . . . . .	114
6.6	Phase transition . . . . .	119
6.7	Influence maximization . . . . .	129
6.8	Discussion . . . . .	135
6.9	Appendix : Contagion on real-world hypergraphs . . . . .	137
6.10	Appendix : Influential groups solutions . . . . .	139
6.11	Supplementary Information . . . . .	140

<b>Conclusion et perspectives</b>	<b>147</b>
<b>Annexes</b>	<b>150</b>
<b>A Simulation des processus de contagion sur réseaux</b>	<b>151</b>
A.1 Avant-propos . . . . .	152
A.2 Résumé . . . . .	153
A.3 Abstract . . . . .	153
A.4 Introduction . . . . .	153
A.5 Markovian spreading processes on networks . . . . .	154
A.6 Exact simulation algorithms . . . . .	155
A.7 Efficiency of the stochastic simulation algorithms . . . . .	161
A.8 Concluding remarks . . . . .	163
A.9 Appendix : Some results for uncorrelated random networks . . . . .	163
A.10 Appendix : Overhead factor due to phantom events . . . . .	164
A.11 Supplemental Material . . . . .	165
<b>Bibliographie</b>	<b>167</b>

# Liste des tableaux

1	Notation asymptotique utilisée dans cette thèse. . . . .	xiii
3.1	Glossaire des symboles importants utilisés aux chapitres 3 et 4. . . . .	48
5.1	Glossaire des symboles importants utilisés au chapitre 5. . . . .	84
6.1	Glossaire des symboles importants utilisés au chapitre 6. . . . .	110
A.1	Glossaire des symboles importants utilisés dans l'annexe A. . . . .	153

# Liste des figures

1.1	Exemple de graphe simple non orienté, le club de karaté de Zachary. . . . .	8
1.2	Hypergraphe et sa représentation en graphe biparti. . . . .	11
1.3	Deux configurations différentes pour un multigraphe dues à la discernabilité des demi-liens. . . . .	15
1.4	Temps de mélange pour la chaîne de Markov associée à l'échantillonnage du modèle des configurations biparti. . . . .	19
1.5	Distribution des degrés pour des graphes extraits de jeux de données. . . . .	21
1.6	Distribution des appartenances et des tailles de groupes pour des graphes bipartis simples extraits de jeux de données. . . . .	23
1.7	Propriétés statistiques des contacts entre individus durant une conférence. . . . .	25
2.1	Évolution temporelle pour les modèles SIR et SIS en champ moyen. . . . .	29
2.2	Transition de phase des modèles SIR et SIS en champ moyen. . . . .	30
2.3	Bistabilité émergente dans le modèle de Granovetter. . . . .	33
2.4	Propriétés de la contagion simpliciale. . . . .	34
2.5	Diagramme des phases du modèle de contagion simpliciale. . . . .	35
2.6	Schéma du formalisme de champ moyen. . . . .	37
2.7	Schéma du formalisme de champ moyen hétérogène. . . . .	38
2.8	Schéma du formalisme des équations maîtresses approximatives centrées sur les noeuds. . . . .	40
2.9	Schéma du formalisme des équations maîtresses approximatives centrées sur les groupes. . . . .	41
2.10	Illustration d'une transition de phase nette et d'une transition de phase étendue. . . . .	43
3.1	Simple illustration of the mesoscopic localization phenomenon. . . . .	50
3.2	Framework for contagions on higher-order networks. . . . .	51
3.3	Comparison of the stationary state near the epidemic threshold in the delocalized and mesoscopic localization regimes. . . . .	56
3.4	Asymptotic localization regimes. . . . .	57
3.5	Impact of finite cutoffs on the boundary separating the localized and delocalized regimes. . . . .	60
3.6	Comparison of the stationary state and the level of localization near the epidemic threshold. . . . .	61
3.7	Mesoscopic localization is imperceptible using the standard inverse participation ratio on nodes. . . . .	62
3.8	Validation of the approximate master equation approach with Monte Carlo simulations. . . . .	67
3.9	Impact of scaling the transmission rate on the boundary separating the localized and delocalized regimes. . . . .	70
3.10	Localized portion of the bifurcation diagram in the mesoscopic localization regime. . . . .	71
4.1	Framework for contagions on higher-order networks. . . . .	74
4.2	Epidemic localization in networks with heterogeneous higher-order structures. . . . .	75
4.3	Local impact of structural interventions in delocalized and localized epidemics. . . . .	78
4.4	Global impact of structural interventions in different localization regimes of SIS and SIR dynamics. . . . .	80
5.1	Modeling contagions and superspreading events through higher-order networks. . . . .	86
5.2	Properties of contagions with nonlinear infection kernels induced by heterogenous exposure. . . . .	89



5.3	Infection kernel resulting from different dose distributions, when considering $\lambda$ to be scale parameter. . . . .	94
5.4	Infection kernel when $\lambda$ is no longer a scale parameter. . . . .	96
5.5	Infection kernel when the condition $1 \ll \tau_c \ll \mathcal{T}$ is no longer respected. . . . .	97
5.6	Nonlinear infection kernels are robust to heterogeneous infectiousness. . . . .	99
5.7	Impact of a fixed mean participation time. . . . .	103
5.8	Temporal evolution and phase diagram for a quenched system. . . . .	105
5.9	Comparison with simulations for the temporal evolution and the phase diagram. . . . .	106
6.1	Mapping of the simplicial contagion model to a hypergraph contagion. . . . .	114
6.2	Fraction $I^*$ of infectious nodes in the stationary state of a contagion model on a hypergraph constructed from high-resolution face-to-face contact data from a primary school in Lyon. . . . .	118
6.3	Fraction $I^*$ of infectious nodes in the stationary state of a contagion model on a hypergraph constructed from coauthorship data in computer science. . . . .	119
6.4	Critical behavior of the function $\mathcal{M}[\rho(r)]$ . . . . .	120
6.5	Phase transition in regular hypergraphs. . . . .	122
6.6	Impact of the third moment of the membership distribution on the phase transition. . . . .	124
6.7	Impact of heterogeneous memberships on the bistability threshold. . . . .	125
6.8	Impact of heterogeneous group sizes on the invasion threshold. . . . .	126
6.9	Mesoscopic localization in large groups. . . . .	127
6.10	Impact of heterogeneous group sizes on the bistability threshold. . . . .	128
6.11	Comparison of the influential spreaders and influential groups strategies. . . . .	133
6.12	Time evolution of the distribution $f_{n_{\max},i}$ when using the random strategy. . . . .	135
6.13	Assessing dynamical correlation around nodes by removing nodes with a large membership. . . . .	141
6.14	Stationary fraction of infected nodes for a threshold model. . . . .	144
6.15	Characterization of the convergence for the randomization procedure. . . . .	145
A.1	Phase transition of the order parameters for two canonical models of spreading. . . . .	155
A.2	Rejection area compared to the acceptance area for the selection of a node with the node-based method. . . . .	157
A.3	Decision tree $\mathcal{T}$ used to select the arrays $\{\mathcal{A}_\beta\}$ . . . . .	159
A.4	Average computation time for a single update using different algorithms. . . . .	162
A.5	Expected overhead factor for the update time due to phantom events. . . . .	165
A.6	Average computation time for a single update using different algorithms. . . . .	166
A.7	Average computation time for a single update for the SIR model using different algorithms. . . . .	166

# Liste des abréviations

## Abréviations françaises

SIS	Susceptible- <b>I</b> nfecté- <b>S</b> usceptible
SIR	Susceptible- <b>I</b> nfecté- <b>R</b> etiré
SEIR	Susceptible- <b>E</b> xposé- <b>I</b> nfecté- <b>R</b> etiré
SIRS	Susceptible- <b>I</b> nfecté- <b>R</b> etiré- <b>S</b> usceptible

Les abréviations sont identiques en anglais, car les termes infecté, exposé et retiré correspondent aux termes *infected*, *exposed* et *recovered* respectivement.

## Abréviations anglaises

HMF	<b>H</b> eterogeneous <b>M</b> ean <b>F</b> ield
AME	<b>A</b> pproximate <b>M</b> aster <b>E</b> quation
QMF	<b>Q</b> uenched <b>M</b> ean <b>F</b> ield
MMCA	<b>M</b> icroscopic <b>M</b> arkov <b>C</b> hain <b>A</b> pproach
MECLE	<b>M</b> icroscopic <b>E</b> pidemic <b>C</b> lique <b>E</b> quation

# Liste des contributions

Les articles suivants furent complétés durant le doctorat. Ceux figurant dans cette thèse sont identifiés à l'aide d'une étoile ★.

- ★ *Influential groups for seeding and sustaining nonlinear contagion in heterogeneous hypergraphs*  
G. St-Onge, I. Iacopini, V. Latora, A. Barrat, G. Petri, A. Allard, L. Hébert-Dufresne  
Commun. Phys. **5**, 25 (2022)
- ★ *Universal Nonlinear Infection Kernel from Heterogeneous Exposure on Higher-Order Networks*  
G. St-Onge, H. Sun, A. Allard, L. Hébert-Dufresne, G. Bianconi  
Phys. Rev. Lett. **127**, 158301 (2021)
- ★ *Social Confinement and Mesoscopic Localization of Epidemics on Networks*  
G. St-Onge, V. Thibeault, A. Allard, L. J. Dubé, L. Hébert-Dufresne  
Phys. Rev. Lett. **126**, 098301 (2021)
- ★ *Master equation analysis of mesoscopic localization in contagion dynamics on higher-order networks*  
G. St-Onge, V. Thibeault, A. Allard, L. J. Dubé, L. Hébert-Dufresne  
Phys. Rev. E **103**, 032301 (2021)
- *Inference, Model Selection, and the Combinatorics of Growing Trees*  
G. T. Cantwell, G. St-Onge, J.-G. Young  
Phys. Rev. Lett. **126**, 038301 (2021)
- *Localization, epidemic transitions, and unpredictability of multistrain epidemics with an underlying genotype network*  
B. J. M. Blake, G. St-Onge, L. Hébert-Dufresne  
PLOS Comput. Biol. **17**, e1008606 (2021)
- *Threefold way to the dimension reduction of dynamics on networks: an application to synchronization*  
V. Thibeault, G. St-Onge, L. J. Dubé, P. Desrosiers  
Phys. Rev. Research **2**, 043215 (2020)
- *Network comparison and the within-ensemble graph distance*  
H. Hartle, B. Klein, S. McCabe, A. Daniels, G. St-Onge, C. Murphy, L. Hébert-Dufresne  
Proc. R. Soc. A **476**, 20190744 (2020)
- *Thresholding normally distributed data creates complex networks*  
G. T. Cantwell, Y. Liu, B. F. Maier, A. C. Schwarze, C. A. Serván, J. Snyder, G. St-Onge  
Phys. Rev. E **101**, 062302 (2020)

- *Detecting structural perturbations from time series with deep learning*  
E. Laurence, C. Murphy, **G. St-Onge**, X. Roy-Pomerleau, V. Thibeault  
arXiv :2006.05232 (2020)
- *Phase transition in the recoverability of network history*  
J.-G. Young, **G. St-Onge**, E. Laurence, C. Murphy, L. Hébert-Dufresne, P. Desrosiers  
Phys. Rev. X **9**, 041056 (2019)
- ★ *Efficient sampling of spreading processes on complex networks using a composition and rejection algorithm*  
**G. St-Onge**, J.-G. Young, L. Hébert-Dufresne, L. J. Dubé  
Comput. Phys. Commun. **240**, 30 (2019)
- *Universality of the stochastic block model*  
J.-G. Young, **G. St-Onge**, P. Desrosiers, L. J. Dubé  
Phys. Rev. E **98**, 032309 (2018)

# Notation

## Développements asymptotiques

Dans cet ouvrage, les développements asymptotiques sont fréquemment utilisés. Dans la limite  $x \rightarrow x_0$ , on caractérise le comportement asymptotique de deux fonctions  $f(x) \geq 0$  et  $g(x) \geq 0$  à l'aide de la notation introduite au tableau 1.

Notation	Signification asymptotique	Définition formelle
$f(x) = o[g(x)]$ ou $f(x) \ll g(x)$	$f(x)$ est dominée par $g(x)$	$\lim_{x \rightarrow x_0} \frac{f(x)}{g(x)} = 0$
$f(x) = O[g(x)]$	$f(x)$ est bornée supérieurement par $g(x)$	$\lim_{x \rightarrow x_0} \frac{f(x)}{g(x)} < \infty$
$f(x) \sim g(x)$	$f(x)$ est proportionnelle à $g(x)$	$\lim_{x \rightarrow x_0} \frac{f(x)}{g(x)} = a$ , $0 < a < \infty$
$f(x) \simeq g(x)$	$f(x)$ tend vers $g(x)$	$\lim_{x \rightarrow x_0} \frac{f(x)}{g(x)} = 1$
$f(x) = \Omega[g(x)]$	$f(x)$ est bornée inférieurement par $g(x)$	$\lim_{x \rightarrow x_0} \frac{f(x)}{g(x)} > 0$
$f(x) = \omega[g(x)]$ ou $f(x) \gg g(x)$	$f(x)$ domine $g(x)$	$\lim_{x \rightarrow x_0} \frac{f(x)}{g(x)} = \infty$

TABLEAU 1 – Notation asymptotique utilisée dans cette thèse.

Remarque : en théorie des nombres, les symboles  $\ll$  et  $\gg$  sont plutôt utilisés pour remplacer  $O$  et  $\Omega$ . Dans cette thèse,  $a \ll b$  implique que  $a$  est beaucoup plus petit que  $b$ , une perspective plus proche de la physique. Pour les développements asymptotiques, les symboles  $\ll$  et  $\gg$  ont comme significations « asymptotiquement dominée par » et « domine asymptotiquement » respectivement, comme dans le livre de Bender et Orszag [24]. Cela permet de comparer plusieurs quantités de manière compacte et naturelle. Par exemple,  $f(x) \ll g(x) \ll h(x)$  dans la limite  $x \rightarrow x_0$  est équivalent à  $f(x) = o[g(x)] \wedge g(x) = o[h(x)]$ .

Aussi, dans la notation de Bachmann–Landau, le symbole  $\sim$  a la signification « tend asymptotiquement vers » (remplacé par le symbole  $\simeq$  dans cette thèse). La notation introduite au tableau 1 permet de distinguer clairement les comportements du type asymptotiquement proportionnel ( $\sim$ ), asymptotiquement égal ( $\simeq$ ) et proportionnel en tout temps ( $\propto$ ).

Les symboles  $O$  et  $\Omega$  sont utilisés pour borner le comportement des fonctions. Par exemple, il serait correct d'écrire  $f(x) = x = O(x) = O(x^2)$  ou  $f(x) = x = \Omega(x) = \Omega(1)$  dans la limite  $x \rightarrow \infty$ . Les relations asymptotiques ne sont pas réversibles en général, par exemple  $o(x) = O(x)$ , mais  $O(x) = o(x)$  est faux [225].

## Systèmes dynamiques

Dans cet ouvrage, on travaille souvent avec des systèmes dynamiques qui atteignent un point fixe. Pour distinguer les variables qui dépendent du temps de leur valeur *stationnaire* dans la limite  $t \rightarrow \infty$ , les points fixes sont toujours distingués par un astérisque (\*). Par exemple, soit  $x(t)$  une variable quelconque, on note  $x(t \rightarrow \infty) \equiv x^*$ .

## Probabilités et processus stochastiques

Pour les besoins de cette thèse, il n'est pas nécessaire de distinguer une variable aléatoire d'une réalisation de cette variable aléatoire. Par conséquent, on note simplement  $P(x)$  la distribution d'une variable aléatoire *discrète*  $x$ ; dans un souci de simplicité, on ne fait pas non plus la distinction entre la distribution  $P$  et la valeur de la fonction de masse évaluée au point  $x$ , soit  $P(x)$ . Similairement, pour un vecteur de probabilité  $(p_i)_{i=1}^N$ , on réfère à  $p_i$  comme étant la « distribution ». Pour une variable aléatoire *continue*  $y$ , on note la densité de probabilités  $\wp(y)$ .

L'espérance d'une quantité  $f(x)$  est notée  $\langle f(x) \rangle$ . Le contexte est suffisant pour clarifier quelle distribution (ou densité de probabilité) est utilisée dans le calcul de l'espérance. Dans certains cas, on note  $\langle f(x)|y \rangle$  si l'espérance est calculée à l'aide de la distribution conditionnelle  $P(x|y)$ .

Pour un processus stochastique, la variable aléatoire  $x(t)$  dépend du temps. Le processus est noté  $(x(t))_{t \in \mathcal{D}_t}$  pour un ensemble  $\mathcal{D}_t$  discret ( $\mathcal{D}_t \subseteq \mathbb{N}_0$ ) ou continu ( $\mathcal{D}_t \subseteq \mathbb{R}^+$ ).

## Vecteurs et matrices

Un vecteur ligne est noté en gras  $\mathbf{x} = [x_1, x_2, \dots, x_N]$ . Une matrice est notée en majuscule et en gras,  $\mathbf{X}$ , et l'élément  $(i, j)$  de cette matrice est noté  $x_{ij}$ . Pour une quantité qui dépend de deux indices, mais qui n'est pas nécessairement associée à une matrice, on utilise la notation  $x_{i,j}$ .

You see, one thing is, I can live with doubt, and uncertainty, and not knowing. I think it's much more interesting to live, not knowing, than to have answers which might be wrong.

---

*Richard Feynman*

# Remerciements

J'ai peine à faire la liste de tous les événements qui ont marqué mon cheminement au doctorat. Les nombreuses rencontres et les nouvelles amitiés, notamment grâce aux nombreux groupes de travail, aux écoles d'été et d'hiver, le séjour de recherche au Vermont et la conférence à Tokyo. La pandémie de COVID-19 qui vient malheureusement tout bouleverser. S'en suit une croissance exponentielle du nombre de projets, pour finalement en arriver à la fin de mon aventure avec l'écriture de cette thèse au pied des Rocheuses. De telles expériences ne seraient pas possibles sans le soutien technique et moral d'une foule de gens. Je prends donc ces quelques lignes pour remercier tous ceux qui, de près ou de loin, ont influencé mon parcours durant ces quatre dernières années.

Mes premiers remerciements sont dédiés au trio de directeurs étoiles qui ont supervisé ma recherche — à tous, je vous tire mon chapeau. D'abord Louis J. Dubé, qui fut mon directeur de recherche à la maîtrise et au début de mon doctorat. Sa sagesse, sa rigueur et sa bienveillance ont teinté ma recherche et m'auront permis de m'épanouir dans ce domaine étrange et fascinant qu'est celui des systèmes complexes.

Ensuite, je tiens à remercier Antoine Allard, qui a repris le flambeau à titre de directeur principal pour mon doctorat. Bien que notre collaboration scientifique soit plus récente, il a laissé sa marque sur mes travaux par une pensée juste et critique. Son écoute et son soutien lors des moments creux, avant même d'avoir le flambeau en main, furent déterminants pour mon parcours. C'est un exemple ponctuel de sa grande dévotion envers les étudiants sous son aile et du succès avec lequel il a su instaurer un climat de travail agréable et stimulant. D'ailleurs, c'est grâce à lui que le groupe Dynamica aura finalement pu voir la lumière du jour, littéralement. Au nom de notre santé mentale à tous, merci mille fois!

Finalement, j'aimerais remercier mon codirecteur de recherche Laurent Hébert-Dufresne, qui fut un réel mentor tout au long de mon doctorat. Je suis le chercheur que je suis aujourd'hui en grande partie grâce à lui, et l'ensemble des travaux inclus dans cette thèse porte sa signature, dans tous les sens du terme. J'admire sa créativité, son esprit débordant et sa vivacité d'esprit; des qualités que j'espère avoir acquises par osmose, notamment durant mon séjour de recherche au Vermont. J'en conserve des souvenirs mémorables, avec Juniper également, trop nombreux pour les énumérer ici — je n'oublierai jamais LE ramen ultime. Même si nous allons certainement nous revoir dans un futur proche en Nouvelle-Angleterre, je vous dis, *Godspeed You!*

Je tiens également à remercier Patrick Desrosiers, Peter Sheridan Dodds et Laetitia Gauvin qui ont accepté de faire partie du jury d'évaluation de cette thèse. Pour certains, les travaux dans cet ouvrage seront plutôt familiers, pour d'autres, ce sera plus éloigner de votre expertise. J'espère que les efforts déployés pour introduire les concepts fondamentaux des processus de contagion vous seront utiles



dans votre lecture. J'aimerais doublement exprimer ma gratitude envers Patrick Desrosiers, pour la prélecture de ce document, mais aussi pour sa présence rassurante, depuis mon premier jour de classe au baccalauréat en physique, jusqu'à l'élaboration de cette thèse. Toujours disponible pour partager son savoir, je le remercie pour sa générosité intellectuelle et son dévouement sans borne.

J'aimerais ensuite remercier tous les membres de la grande famille Dynamica. D'abord, un merci spécial à Jean-Gabriel Young, avec qui j'ai énormément appris sur l'inférence, la conception d'algorithmes et bien plus, lors de nos nombreux échanges pendant mon parcours aux cycles supérieurs. Son influence s'est manifestée sous plusieurs rôles : en tant que collègue, en tant que *lifting bro*, mais c'est certainement comme mentor qu'il aura laissé sa plus grande marque. Je tiens ensuite à remercier Charles, Edward, Vincent et Xavier, fiers membres du sous-groupe Dynamitas, avec qui j'aurai passé des moments inoubliables. L'humour pince-sans-rire unique, les discussions philosophiques profondes et le basket-ball improvisé resteront gravés dans ma mémoire — ma cheville, elle aussi, s'en souvient très bien. Je tiens aussi à saluer Jean-Luc, dernier représentant de la division lumière : le groupe de recherche ne produira plus jamais de figures aussi majestueuses. Merci également à Béatrice, pour ton esprit d'initiative remarquable, et dont la simple présence inspire une joie de vivre *contagieuse*.

Puisque la famille Dynamica s'est agrandie à un rythme effréné durant les dernières années, il m'est impossible de tous vous remercier personnellement. Bien que je n'aie pas eu la chance d'échanger avec la relève autant que je l'aurais souhaité, je vous remercie pour ces discussions enrichissantes et je vous souhaite à tous un agréable parcours au sein du groupe.

En dehors du cercle Dynamica, je salue l'ensemble des membres du *Laboratory for Structure and Dynamics* (LSD) au Vermont, cette famille éloignée que j'ai eu la chance d'apprendre à connaître lors de mon passage dans la contrée ; les gens incroyables du *Network Science Institute* rencontrés lors de l'école d'hiver sur les réseaux complexes ; et les personnes fantastiques avec qui j'ai passé un mois à Santa Fe lors de l'école d'été sur les systèmes complexes. Je tiens à remercier mes amis, et plus particulièrement ceux qui ont emprunté un parcours semblable au mien : Jean-Samuel, coéquipier éternel, ainsi que Shanny, Maxime et Félix. Entreprendre des études aux cycles supérieurs, c'est aussi accepté d'être incompris de la masse ; il est toujours réconfortant de pouvoir partager ses problèmes et ses victoires avec des gens qui en apprécient la valeur.

J'aimerais enfin remercier mes parents, Yves et Josée, pour votre appui et votre fierté indéfectibles ; mon frère Olivier, pour me ramener sur terre de temps à autre ; Denise et Nicole, pour votre générosité sans borne et vos encouragements — ce n'est pas ma grand-mère qui va m'arrêter dans mon élan ; et Anthony, ami de longue date qui me connaît mieux que quiconque. Si on se fie à Platon, la philosophie est directement liée au dialogue : je te dois alors une grande partie de la sagesse qui m'habite aujourd'hui.

Et un dernier remerciement à ma douce Marie-Ange. Malgré les épreuves, son soutien moral transcende les distances et m'a apporté l'énergie nécessaire à persévérer dans cette longue aventure.

# Avant-propos

Les articles qui suivent sont directement incorporés à la thèse. Le contenu n'en a été modifié que pour se conformer au format exigé par la Faculté des études supérieures et postdoctorales de l'Université Laval. La seule exception est le chapitre 5, pour lequel la notation et la terminologie ont été adaptées par souci de cohérence avec les autres chapitres. À titre de premier auteur, j'ai été celui qui a principalement contribué à l'élaboration des projets, aux développements théoriques, à la réalisation des calculs numériques et à la rédaction des manuscrits. Les autres auteurs ont apporté leur soutien à chacun des aspects susmentionnés, surtout à l'élaboration des projets et à la révision des manuscrits auxquels ils ont participé.

- CHAPITRE 3  
*Master equation analysis of mesoscopic localization in contagion dynamics on higher-order networks*  
**G. St-Onge**, V. Thibeault, A. Allard, L. J. Dubé, L. Hébert-Dufresne  
Phys. Rev. E **103**, 032301 (publication : mars 2021)
- CHAPITRE 4  
*Social Confinement and Mesoscopic Localization of Epidemics on Networks*  
**G. St-Onge**, V. Thibeault, A. Allard, L. J. Dubé, L. Hébert-Dufresne  
Phys. Rev. Lett. **126**, 098301 (publication : mars 2021)
- CHAPITRE 5  
*Universal Nonlinear Infection Kernel from Heterogeneous Exposure on Higher-Order Networks*  
**G. St-Onge**, H. Sun, A. Allard, L. Hébert-Dufresne, G. Bianconi  
Phys. Rev. Lett. **127**, 158301 (publication : octobre 2021)
- CHAPITRE 6  
*Influential groups for seeding and sustaining nonlinear contagion in heterogeneous hypergraphs*  
**G. St-Onge**, I. Iacopini, V. Latora, A. Barrat, G. Petri, A. Allard, L. Hébert-Dufresne  
Commun. Phys. **5**, 25 (publication : janvier 2022)
- ANNEXE A  
*Efficient sampling of spreading processes on complex networks using a composition and rejection algorithm*  
**G. St-Onge**, J.-G. Young, L. Hébert-Dufresne, L. J. Dubé  
Comput. Phys. Commun. **240**, 30 (publication : juillet 2019)

# Introduction

## Épidémies, rumeurs et désinformation

Le monde est aux prises en ce moment avec deux types de pandémies. L'une est causée par la propagation du virus SARS-CoV-2, lequel est responsable de la COVID-19, une maladie aéroportée possiblement mortelle dont les symptômes sont multiples, le plus grave étant la détresse respiratoire. L'autre est d'une tout autre nature : la désinformation reliée à la COVID-19 se propageant à une vitesse fulgurante, surtout grâce aux médias sociaux. Les symptômes de cette *contagion sociale* sont l'adhésion à des thèses conspirationnistes, la négation des bienfaits des vaccins contre la COVID-19, etc. Malheureusement, les deux ont des conséquences néfastes sur notre société.

Malgré l'importante différence au niveau des mécanismes de transmission, ces deux pandémies résultent des *interactions* entre les individus d'une population et sont modélisées d'une manière analogue. Les *processus de contagion* décrivent l'évolution des phénomènes biologiques, sociaux ou même technologiques que l'on associe à quelque chose qui se propage au sein d'une population. Pensons par exemple à un agent pathogène (virus, bactérie, parasite, etc.), une rumeur [66], une innovation [244], un mouvement social [112] ou un virus informatique [233]. En dépit du caractère général de la propagation, on utilise dans cette thèse la terminologie propre à l'épidémiologie ; un individu *infecté* peut aussi bien s'interpréter comme un individu ayant adopté un comportement social et qui, par ses actions, favorise l'adoption de ce comportement par d'autres individus.

La perspective que l'on utilise pour étudier les processus de contagion est principalement fondée sur la physique statistique hors d'équilibre. En effet, l'hypothèse de base est celle d'un système *probabiliste*, dans lequel l'exposition à un certain nombre d'individus infectés est associée à une probabilité d'infection. Même si les interactions sociales ne sont pas fondamentalement aléatoires<sup>1</sup>, une modélisation purement statistique du système devient justifiée lorsque la population est suffisamment grande, de manière analogue à la théorie cinétique des gaz. De plus, le cadre théorique des processus de contagion est semblable à celui des systèmes de *spins*, comme les modèles d'Ising ou de Potts — chaque individu possède un état : *susceptible* à la maladie ou infecté. En variant certains paramètres, comme le taux d'infection par exemple, on observe également une *transition de phase*. Typiquement, dans l'une de ces phases, la maladie n'est pas en mesure d'envahir la population au-delà d'une poignée d'individus, tandis que dans l'autre, l'épidémie atteint une taille *macroscopique* (une fraction importante de la population est affectée par la maladie).

D'ailleurs, les modèles standards en épidémiologie supposent qu'une population d'individus se comporte à l'instar d'un gaz de particules : chaque individu peut entrer en contact avec n'importe quel autre

---

1. On se permet ici de négliger les effets quantiques.

individu de manière équiprobable. La population est homogène et tous les individus d'un même état sont indiscernables, ce qui permet de limiter considérablement le nombre d'équations pour décrire le système. Cette hypothèse est souvent nommée la *loi d'action de masse* ou, en physique statistique, l'approximation de *champ moyen*. Il est possible d'ajouter du réalisme en subdivisant la population en « compartiments », selon différents facteurs : âge [206], statut socioéconomique, etc. ; cette approche porte le nom de *formalisme compartimental*.

Toutefois, la société n'est pas un gaz, pour lequel les particules remplissent un volume de manière complètement désordonné et homogène. La société n'est pas non plus un cristal, dont les atomes interagissent de manière parfaitement ordonnée et symétrique. Non, la structure des interactions sociales se trouve quelque part entre les deux, dans cette étendue vaste que l'on qualifie de *complexe*.

## Complexité et science des réseaux

Bien qu'il n'existe pas de définition universelle d'un *système complexe*, plusieurs s'entendent pour dire qu'il s'agit d'un système composé de plusieurs éléments en interaction qui possède des propriétés *émergentes*, lesquelles ne peuvent pas être prédites uniquement à partir des éléments seuls. Comme propriété émergente, mentionnons la synchronisation des neurones, l'aimantation spontanée d'un système de spins ou, dans le cadre de cette thèse, l'émergence d'une épidémie. D'autres propriétés sont généralement ajoutées à cette définition des systèmes complexes : adaptation, structure et hiérarchie sont parmi les caractéristiques les plus courantes. Dans un texte fondateur [291], Warren Weaver distingua les problèmes de complexité *désorganisée* — comme un gaz — des problèmes de complexité *organisée*, dans lesquels les interactions entre les éléments possèdent une structure complexe, ni parfaitement régulière, ni complètement aléatoire.

La *science des réseaux* [220] s'intéresse à l'étude de la structure de ces systèmes complexes organisés et de son incidence sur la dynamique. Pour ce faire, les éléments du système sont associés à des noeuds et les interactions entre les éléments sont représentées par des liens. Ce paradigme permet de comparer des systèmes de nature différente et de modéliser divers processus dynamiques partageant une même structure. Entre autres, la science des réseaux offre une avenue pour répondre aux questions suivantes : quelles propriétés structurelles universelles sous-tendent les phénomènes naturels et notre société ? Quels mécanismes d'évolution et d'auto-organisation des réseaux mènent à ces propriétés ? Et quel est leur effet sur les processus dynamiques ? Parmi les propriétés universelles des réseaux, mentionnons l'*hétérogénéité* de la structure, comme le fait que certains noeuds soient beaucoup plus connectés que d'autres au reste du système [16], ce qui influence considérablement la dynamique de ces systèmes [19].

C'est sur ce terreau fertile que prend racine le domaine de l'épidémiologie sur réseaux [163, 231]. Puisque la structure des contacts n'est jamais parfaitement connue, le concept de *réseau aléatoire* est devenu un outil essentiel pour étudier de manière rigoureuse l'incidence des différentes propriétés structurelles sur les contagions. Un réseau aléatoire permet de fixer un certain nombre de propriétés par contraintes tout en maximisant l'*entropie*, minimisant par le fait même l'information injectée dans le système [145]. Il est alors possible de généraliser de manière systématique et rigoureuse les formalismes compartimentaux aux processus de contagion sur réseaux aléatoires — les caractéristiques structurelles fixées font alors partie intégrante des attributs distinguant les différents compartiments.

## Processus de contagion sur réseaux d'ordre supérieur

Au cours des deux dernières décennies, la dynamique des systèmes complexes fut surtout étudiée en considérant des réseaux d'interactions *dyadiques* (impliquant une paire d'éléments), ou plus formellement, des *graphes*. L'hypothèse sous-jacente est que la dynamique dépend uniquement des dyades. Par exemple, une supposition typique en épidémiologie est que la transmission d'une maladie s'effectue via une paire d'individus : l'un susceptible, l'autre infecté. Que cette interaction dyadique découle d'un événement impliquant plusieurs individus n'a donc pas d'incidence sur l'évolution de la dynamique selon cette hypothèse. Le graphe des systèmes complexes possède souvent une structure *communautaire* ou *modulaire* [103] caractérisée par une densité d'interactions plus élevée à l'intérieur de certains *groupes* de noeuds. Ces groupes peuvent correspondre à un cercle d'amis dans un réseau social ou à une région du cerveau possédant une fonction particulière. Une structure modulaire a une grande influence sur la dynamique des systèmes complexes [73, 104, 125, 215, 293] et peut même mener à l'émergence d'états exotiques, comme les *chimères* dans les dynamiques de synchronisation [275]. Bien que ces modules aient un effet sur l'évolution temporelle d'un processus dynamique sur réseaux, ils n'ont aucun effet sur la dynamique en elle-même (c'est-à-dire, les règles la gouvernant restent inchangées).

Plusieurs travaux récents contestent cette hypothèse selon laquelle les interactions dyadiques déterminent entièrement l'évolution des systèmes complexes. La tendance actuelle est plutôt de considérer que les systèmes complexes sont, la plupart du temps, forgés par des interactions d'*ordre supérieur* (impliquant un groupe arbitraire d'éléments), et qu'il est impossible de décomposer ces interactions de groupe en dyades sans en perdre l'essence. Par exemple, une discussion de groupe est très différente de plusieurs interactions séparées ou, de manière analogue, la collaboration d'une équipe sur un projet n'est pas la somme des interactions dyadiques. Il fut également montré que les interactions d'ordre supérieur stabilisent les dynamiques de compétition, ce qui aide à expliquer la robustesse de la coexistence des espèces en écologie [113]. Ces exemples mettent en évidence le caractère fondamental des interactions de groupe. C'est pourquoi la science des réseaux se tourne de plus en plus vers les représentations d'ordre supérieur, comme les *hypergraphes* et les *complexes simpliciaux* [22].

L'utilisation des réseaux d'ordre supérieur n'est pas nouvelle. Par exemple, l'un des articles fondateurs de la science des réseaux [221], par Mark Newman, Steven H. Strogatz et Duncan J. Watts, traite des graphes bipartis, lesquels peuvent être reliés aux hypergraphes. Plusieurs travaux ont exploité cette représentation de manière implicite [6, 7, 125, 218], mais les interactions de groupe étaient généralement un prétexte pour introduire une structure modulaire ou des motifs particuliers dans les graphes. Récemment, le rôle des interactions de groupe a été considéré de manière plus explicite dans l'évolution des processus de contagion [36]. L'avenue de recherche fut notamment popularisée par un modèle de contagion sociale sur complexes simpliciaux [142], dans lequel les interactions d'ordre supérieur sont utilisées pour induire une pression sociale. Plusieurs autres travaux ont suivi [40, 59, 88, 147, 196], améliorant la description analytique et poussant plus loin l'analyse théorique.

C'est dans cette mouvance que s'inscrivent les contributions originales présentées dans cette thèse. Notre objectif est de développer des approches théoriques qui permettent de modéliser adéquatement la phénoménologie des processus de contagion sur réseaux d'ordre supérieur. Nous mettons l'accent sur les formalismes compartimentaux, grâce auxquels nous sommes en mesure de démontrer plusieurs résultats analytiques illustrant le rôle central des interactions de groupe sur la dynamique de propagation.

## Organisation de la thèse

Cette thèse est divisée en trois parties, qui contiennent chacune deux chapitres. La partie I introduit les concepts de bases pour l'étude des processus de contagion sur réseaux. Les parties II et III exposent nos contributions originales. Ce sont principalement des articles publiés (ou en voie de l'être) dans des revues scientifiques. La terminologie et la notation employées dans les articles ont été harmonisées avec le reste de la thèse.

Les approches analytiques sont mises de l'avant pour l'étude des processus de contagion ; les considérations de nature numérique sont reléguées à l'annexe A, comprenant un article dans lequel nous démontrons comment accélérer de manière substantielle la simulation des processus de contagion.

**Partie I** – Au chapitre 1, on présente formellement les graphes, les hypergraphes, la théorie des réseaux aléatoires ainsi que les propriétés statistiques typiques des systèmes sociaux. Au chapitre 2, on introduit la théorie des processus stochastiques de contagion. On illustre les concepts de contagion *simple* et de contagion *complexe* à l'aide de quelques modèles faciles à comprendre, suivis de certains des principaux formalismes compartimentaux développés pour l'étude des contagions sur réseaux. Finalement, on discute de phénomènes exotiques importants dans le cadre de cette thèse, soit la *localisation* et les transitions de phase *étendues*.

**Partie II** – Au chapitre 3, nous étudions la dynamique Susceptible-Infecté-Susceptible (SIS) grâce à un formalisme d'équations maîtresses approximatives centrées sur les groupes. Nous démontrons l'émergence d'un phénomène de *localisation mésoscopique* : pour certaines structures hétérogènes, seuls les groupes de grande taille soutiennent le processus de contagion près du point critique de la transition de phase. Au chapitre 4, nous montrons que ce phénomène a une incidence importante sur l'effet des mesures de contrôle visant à limiter les regroupements au-delà d'une certaine taille, à l'instar de ce qui fut instauré pour endiguer la pandémie de COVID-19. Ces deux chapitres sont basés sur deux articles publiés conjointement.

**Partie III** – Au chapitre 5, nous étudions un modèle dans lequel les individus développent la maladie uniquement s'ils reçoivent une *dose infectieuse minimale* à l'intérieur d'un groupe. En couplant cet effet avec des temps d'exposition hétérogènes, nous démontrons que les individus deviennent infectés selon une probabilité d'infection non linéaire *universelle* en loi de puissance (en fonction du nombre d'individus infectés dans le groupe). Selon l'exposant de la loi de puissance, la taille de l'épidémie peut avoir une croissance super-exponentielle et la transition de phase peut devenir discontinue avec un régime bistable. Au chapitre 6, nous poussons plus loin l'analyse des contagions non linéaires sur réseaux d'ordre supérieur grâce aux équations maîtresses approximatives. Nous caractérisons l'incidence de la localisation mésoscopique sur les propriétés critiques du système, comme l'inhibition de la bistabilité. Nous montrons également que les groupes peuvent être plus *influent*s que les individus ultra-connectés d'une population lorsque la contagion est suffisamment non linéaire.

## Première partie

# Notions de base pour les processus de contagions sur réseaux

# Chapitre 1

## Représentation en réseau des systèmes complexes

Un réseau est une abstraction mathématique encodant la structure d'un système composé d'éléments en interaction. Les éléments de base sont représentés par des noeuds, et les interactions entre ceux-ci par des liens. Le noeud et le lien peuvent avoir différentes significations, ce qui en fait un outil extrêmement flexible pour représenter divers types de systèmes : les interactions protéine-protéine [146], un écosystème [191], les interactions sociales dans une communauté [241], etc.

Puisque dans cette thèse on s'intéresse aux processus de contagion, on met l'accent sur la représentation en réseaux des interactions *sociales*. Les noeuds représentent des individus et les liens peuvent représenter différents types de relations [220] : amitié, relation professionnelle, relation romantique ou sexuelle, etc. On s'intéresse particulièrement aux réseaux sociaux qui sont informatifs pour prédire les *contacts* potentiels entre individus — un aspect important dans le cadre des contagions biologiques — ou pouvant servir à la transmission d'information, nécessaire aux processus de contagion sociale. Notons que les réseaux d'*information* [220] (Internet, réseaux de pair-à-pair, etc.) peuvent aussi servir de substrat à la propagation, comme pour la transmission d'un virus informatique ou d'un maliciel entre serveurs.

Dans ce chapitre, on commence par une description mathématique des représentations en réseaux qui seront utiles pour les chapitres suivants. Bien que le coeur de cette thèse concerne les processus de contagion sur réseaux d'ordre supérieur, il est plus naturel et pédagogique d'introduire d'abord les réseaux d'interactions dyadiques — les *graphes*. De plus, la plupart des formalismes analytiques développés dans les dernières années pour caractériser les processus de contagion utilisent cette représentation (voir chapitre 2). On passe ensuite aux réseaux d'ordre supérieur, plus spécifiquement aux *hypergraphes* et à leur représentation en *graphe biparti*, et à la théorie des réseaux aléatoires. Finalement, on discute de certaines des propriétés universelles des réseaux sociaux, lesquelles serviront de guide aux chapitres suivants pour l'étude des processus de contagion.

Notons qu'il existe différentes manières d'incorporer du réalisme à un réseau [220], notamment à l'aide d'une structure à niveaux multiples [165], ou en utilisant un réseau temporel [134, 135] ou adaptatif [115, 231]. On se limite toutefois au cas plus simple d'une structure statique avec une seule couche afin de pousser l'analyse théorique des processus dynamiques.



Autre remarque : tout au long de ce document, le terme réseau possède une connotation générale et englobe les différents types de représentations utilisant les noeuds et les liens comme unité de base ; un vocabulaire plus précis (graphe, hypergraphe, etc.) sera utilisé de manière ponctuelle afin de clarifier l’objet mathématique à l’étude.

## 1.1 Graphe : interactions dyadiques

Dans cette thèse, on utilise le terme *graphe* de manière générale pour englober les différents types de réseaux dyadiques. Peu importe la représentation, les éléments du système sont toujours associés à un ensemble de noeuds  $\mathcal{V} = \{1, 2, \dots, N\}$ , où  $N$  est le nombre total de noeuds.

Il y a trois aspects principaux qui distinguent les types de graphes : si les liens sont *orientés*, si les *boucles* sont permises et si les liens multiples sont permis. Un lien orienté partant du noeud  $i$  vers le noeud  $j$  est noté  $(i, j)$ , alors qu’un lien non orienté est noté  $\{i, j\}$ . Une boucle est un lien reliant un noeud à lui-même et peut être notée  $(i, i)$  ou  $\{i\}$  selon le cas.

On se concentre sur les graphes *non orientés*, permettant de représenter les interactions réciproques entre les éléments d’un système. Pour les structures sociales, c’est typiquement le cas<sup>1</sup> : par exemple, si les liens d’un graphe représentent les relations d’amitié entre des individus, on s’attend à ce que les liens soient réciproques. De manière semblable, un contact physique entre deux individus est généralement considéré comme non orienté.

Le type le plus contraint est le *graphe simple*  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ , où  $\mathcal{E} \subseteq \{\{i, j\} | (i, j) \in \mathcal{V} \times \mathcal{V} \wedge i \neq j\}$  est l’ensemble des liens. La plupart des réseaux sociaux respectent ces contraintes : un individu ne peut pas être ami avec lui-même ou être deux fois l’ami d’un autre individu, et la même logique s’applique pour plusieurs types de relations entre individus (collègues de travail, coéquipiers dans un sport, partenaires sexuels, etc.). Par exemple, la figure 1.1 présente le graphe du fameux club de karaté de Zachary, dans lequel un lien indique que les individus interagissaient à l’extérieur du club.

Un autre objet auquel on s’intéresse est le *multigraphe avec boucles*  $\mathcal{G} = (\mathcal{V}, \tilde{\mathcal{E}})$ , où  $\tilde{\mathcal{E}} = (\mathcal{E}, \psi)$  est le multiensemble des liens,  $\mathcal{E} \subseteq \{\{i, j\} | (i, j) \in \mathcal{V} \times \mathcal{V}\}$  est l’ensemble des liens et  $\psi : \mathcal{E} \rightarrow \mathbb{N}$  donne la multiplicité des liens de l’ensemble  $\mathcal{E}$ . Par exemple, s’il y a deux liens entre le noeud  $i$  et le noeud  $j$ , alors  $\{i, j\} \in \mathcal{E}$  et  $\psi(\{i, j\}) = 2$ . Bien qu’un multigraphe avec boucles soit une représentation moins commune des interactions sociales, c’est un objet mathématique qui intervient souvent, notamment dans les modèles de réseaux aléatoires. Parfois, il est plus simple d’accepter les boucles et les liens multiples que de forcer un graphe simple, le tout sans conséquence majeure sur les propriétés dynamiques du système (voir section 1.3).

Voici quelques concepts fondamentaux reliés aux graphes.

**Composante** – Un graphe peut contenir plusieurs composantes disjointes  $\mathcal{V}_\alpha \subseteq \mathcal{V}$  associées à des sous-graphes isolés, mais connexes. Tous les noeuds appartenant à une même composante peuvent être rejoints par les liens du réseau. À l’inverse, il est impossible de rejoindre deux noeuds appartenant à des composantes séparées. Pour l’étude des systèmes dynamiques sur graphes, on s’intéresse surtout

---

1. Bien que la relation que l’on tente de mesurer soit souvent de nature non orientée, les données recueillies sont parfois asymétriques et mieux représentées par un graphe orienté (voir le chapitre 4 du livre de Mark Newman [220]). On s’intéresse donc à la version non orientée qu’il est possible de reconstruire à partir des données.

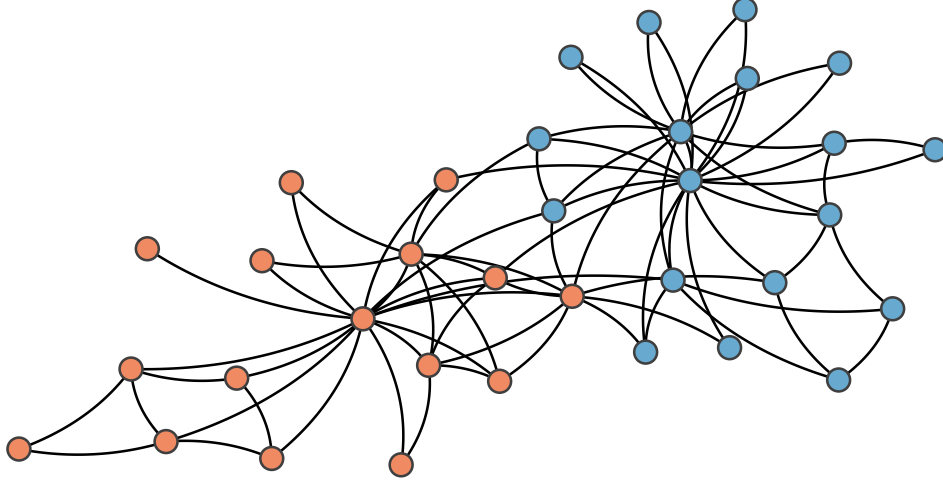


FIGURE 1.1 – Exemple de graphe simple non orienté, le club de karaté de Zachary [303] (nom de l’auteur de l’article associé). Les noeuds représentent des individus qui appartenaient au club et un lien indique que les deux individus interagissaient à l’extérieur du club. Ce réseau emblématique fut utilisé abondamment pour illustrer le concept de *communauté* en science des réseaux à la suite de son apparition dans un article de Michelle Girvan et Mark Newman [103]. Les noeuds rouges et bleus représentent deux factions associées aux individus ayant suivi l’instructeur ou l’administrateur du club respectivement après une dispute entre ces deux derniers. Les méthodes de détection de communautés sur graphes parviennent généralement à distinguer ces deux sous-groupes à partir de la structure du graphe uniquement.

aux graphes contenant une seule composante, car deux composantes séparées agissent généralement comme deux systèmes indépendants.

**Voisinage** – Le voisinage d’un noeud  $i$ , que l’on note  $\partial i$ , correspond au multiensemble de ses voisins, c’est-à-dire les noeuds  $j$  connectés à  $i$  par un lien  $\{i, j\} \in \mathcal{E}$  et tenant compte de leur multiplicité le cas échéant. Si le noeud  $i$  possède une boucle, alors  $\partial i$  contient deux fois l’indice  $i$ , car il est son propre voisin par les deux extrémités du lien.

**Matrice d’adjacence** – On peut représenter la structure complète d’un graphe à l’aide de différentes matrices [220]. La plus commune est la matrice d’adjacence  $\mathbf{A}$  de dimension  $N \times N$  où l’élément  $a_{ij}$  indique le nombre de liens partant du noeud  $j$  vers le noeud  $i$ . Si le noeud  $i$  possède une boucle, alors  $a_{ii} = 2$ , car il y a deux liens sortants du noeud  $i$  — les deux extrémités de la boucle. Il est donc souvent plus intuitif d’interpréter  $a_{ij}$  comme le nombre de *demi-liens* partant de  $j$  vers  $i$ . Pour les graphes non orientés, la matrice d’adjacence est symétrique ( $a_{ij} = a_{ji}$ ) et la somme de tous les éléments,

$$\sum_{i,j=1}^N a_{ij} = 2M, \quad (1.1)$$

correspond au double du nombre de liens du graphe,  $M$ .

**Degré** – Le degré  $k_i$  du noeud  $i$  est le nombre de voisins qu’il possède, donc  $k_i = |\partial i|$ . Le degré d’un noeud se calcule également à l’aide de la matrice d’adjacence

$$k_i = \sum_{j=1}^N a_{ij}. \quad (1.2)$$

Par conséquent, une définition équivalente du degré  $k_i$  est le nombre de demi-liens du noeud  $i$ . Le degré est une propriété fondamentale, mesurant à quel point un chaque noeud est connecté au reste du système. Malgré sa simplicité, le degré des noeuds détermine en grande partie l'état des dynamiques sur graphes (voir chapitre 2). On définit par le fait même la suite des degrés  $(k_i)_{i=1}^N$ , que l'on exprime aussi sous la forme vectorielle  $\mathbf{k} = [k_1, k_2, \dots, k_N]^\top$ . On note également  $k_{\min}$  la valeur minimale et  $k_{\max}$  la valeur maximale de la suite des degrés.

**Distribution des degrés** – La suite des degrés encapsule toute l'information sur le degré des noeuds, mais elle est peu pratique pour caractériser ou comparer deux graphes. On préfère souvent synthétiser l'information à l'aide de la distribution des degrés  $P(k) \equiv p_k$ , la probabilité qu'un noeud choisi aléatoirement (et uniformément) soit de degré  $k$ . C'est un concept fort utile et qui suggère naturellement une interprétation probabiliste pour la structure. La distribution respecte la condition de normalisation

$$\sum_{k=k_{\min}}^{k_{\max}} p_k = 1 . \quad (1.3)$$

On définit également les moments d'ordre  $l$  comme suit :

$$\langle k^l \rangle = \sum_k k^l p_k , \quad (1.4)$$

où il est sous-entendu que l'on somme sur l'entièreté du support.

### 1.1.1 Agrégation et structure mésoscopique des graphes

La distribution des degrés offre un portrait global de l'échelle *microscopique* du graphe, c'est-à-dire au niveau des noeuds et de leur voisinage. Par contre, deux graphes avec une même distribution des degrés peuvent avoir des structures très différentes. La prochaine question que l'on peut se poser est à quel point les voisins d'un noeud partagent des liens entre eux, autrement dit, à quel point les noeuds voisins ont des voisins en commun. Pour un graphe simple, on mesure cette tendance grâce au *coefficient d'agrégation*

$$C = \frac{6N_{\Delta}}{N_{\wedge}} , \quad (1.5)$$

où  $N_{\Delta}$  est le nombre de triangles, c'est-à-dire le nombre de triplets  $\{i, j, \ell\}$  complètement connectés, et  $N_{\wedge}$  est le nombre de parcours de longueurs deux, donc le nombre de couples ordonnés  $(i, j, \ell)$  pour lesquels les liens  $\{i, j\}$  et  $\{j, \ell\}$  existent. Le facteur six s'explique par le fait qu'il y a six parcours de longueur deux possibles dans un triangle.

Les graphes réels — et particulièrement les réseaux sociaux — ont un coefficient d'agrégation élevé par rapport à ce que l'on pourrait s'attendre à partir de leur distribution des degrés seulement [220] (voir section 1.3). Ceci est dû, entre autres, à la présence de *communautés* [103, 220]. Par exemple, les gens d'un réseau social sont amis souvent non pas par chance, mais par contacts à travers un regroupement : équipe sportive, groupe de danse, classe scolaire, etc. Ces communautés favorisent les relations, ce qui renforce le nombre de liens à l'intérieur d'un groupe de noeuds et explique alors le haut coefficient d'agrégation. Le club de karaté de Zachary à la figure 1.1 est populaire, notamment à cause de la séparation du club en deux communautés, lesquelles peuvent être *inférées* simplement à partir du graphe des interactions [103].

Plus généralement, un haut coefficient d'agrégation peut s'expliquer par une structure complexe à l'échelle *mésoscopique* [301]. Pour bien aborder le sujet, il est important que l'on définisse clairement les échelles d'un système : tout ce qui se rattache à un noeud et à son voisinage direct relève de l'échelle *microscopique* ; les quantités se rapportant à l'ensemble des noeuds d'un réseau relèvent de l'échelle *macroscopique*. L'échelle *mésoscopique* quant à elle touche à tout ce qui est entre les deux, c'est-à-dire dès que l'on considère un *groupe* de noeuds  $\mathcal{B} \subset \mathcal{V}$ . Le groupe de noeuds  $\mathcal{B}$  pourrait avoir une taille qui dépend du nombre de noeuds du réseau,  $N$ , ou une taille fixe indépendante de  $N$ .

Par conséquent, quand on parle d'une structure à l'échelle mésoscopique, on s'intéresse généralement à des groupes de noeuds  $\mathcal{B}_1, \mathcal{B}_2, \dots, \mathcal{B}_q$  et à leurs interactions. Ces interactions sont mises de l'avant notamment dans le *modèle stochastique par blocs* [4], un modèle de réseau aléatoire dans lequel on assigne une probabilité de connexion  $p_{r,s}$  entre chaque paire de noeuds provenant des blocs  $\mathcal{B}_r$  et  $\mathcal{B}_s$ <sup>2</sup>. Avec différents choix pour  $\{p_{r,s}\}_{r,s=1}^q$ , il est possible de créer des graphes avec une structure communautaire, une structure *coeur-périphérie*<sup>3</sup> et bien d'autres. À l'instar du degré des noeuds, cette structure complexe à l'échelle mésoscopique affecte grandement les processus dynamiques sur réseaux [5, 73, 104, 125].

## 1.2 Réseaux d'ordre supérieur

Le graphe met l'accent sur les interactions dyadiques ou par paires de noeuds. Cette représentation est généralement suffisante si la dynamique du système se décrit adéquatement à l'échelle microscopique. Toutefois, certains systèmes complexes possèdent une structure et une dynamique qui opère également à l'échelle mésoscopique.

Par exemple, la perspective standard est que les contagions de nature *biologique* — la transmission d'un virus ou d'une bactérie entre individus — s'effectuent entre deux personnes, l'une *infectée* et l'autre *susceptible* à l'agent pathogène. Selon cette perspective, le graphe serait une représentation suffisante du réseau social permettant la transmission. Dans cette thèse, nous remettons en question cette perspective classique, notamment au chapitre 5.

Pour les contagions de nature *sociale* — la formation d'opinion [289], la diffusion des innovations [244], etc. —, le rôle important des groupes d'individus sur la dynamique est un aspect de plus en plus étudié [142]. Par exemple, soit un individu possédant des amis appartenant à plusieurs groupes distincts (équipe sportive, environnement de travail, etc.). L'intuition suggère qu'une opinion partagée par un sous-ensemble de ses amis sera perçue différemment s'ils sont tous à l'intérieur du même groupe, par rapport à s'ils sont distribués uniformément dans les groupes auxquels il appartient. Entre autres, l'individu risque de rencontrer les gens associés à un même groupe de manière simultanée, auquel cas la pression sociale affecte inévitablement la dynamique.

Cela motive une description mathématique de la structure des systèmes complexes qui considère explicitement les interactions d'*ordre supérieur*, impliquant un nombre  $n > 2$  d'individus — les réseaux d'ordre supérieur. Il existe plusieurs types de représentation possibles [22, 277], dont les *complexes*

2. Notons qu'il existe plusieurs variantes du modèle stochastique par blocs, par exemple une version permettant d'inclure l'hétérogénéité du degré des noeuds [150].

3. Un coeur de noeuds fortement connectés entre eux avec des noeuds en périphérie rattachés au coeur, mais faiblement connectés entre eux.

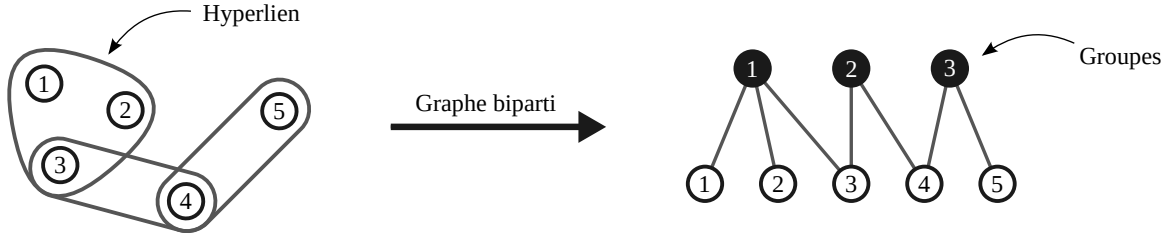


FIGURE 1.2 – Hypergraphe et sa représentation en graphe biparti. Les hyperliens  $\{1, 2, 3\}$ ,  $\{3, 4\}$  et  $\{4, 5\}$  deviennent les groupes 1, 2 et 3.

*simpliciaux*<sup>4</sup>, mais la plus flexible et appropriée pour les travaux de cette thèse est l'*hypergraphe*. Les interactions de groupe sont alors représentées par des *hyperliens*  $e \subseteq \mathcal{V}$  permettant d'inclure un nombre arbitraire de noeuds.

Un hypergraphe peut aussi avoir des boucles, c'est-à-dire des hyperliens de la forme  $e = \{i\}$  contenant un seul noeud. Un hypergraphe simple n'a pas d'hyperliens multiples ni de boucles ; dans le cas contraire, on parlera d'un hypergraphe multiple, avec ou sans boucles. Comme pour les graphes, on se concentre sur les hypergraphes non orientés.

Dans sa forme la plus générale, il y a l'hypergraphe multiple avec boucles  $\mathcal{H} = (\mathcal{V}, \tilde{\mathcal{E}})$ , où  $\tilde{\mathcal{E}} = (\mathcal{E}, \psi)$  est un multiensemble d'hyperliens,  $\mathcal{E} \subseteq \mathcal{P}(\mathcal{V}) \setminus \emptyset$  est l'ensemble des hyperliens possibles,  $\mathcal{P}(\mathcal{V})$  est l'ensemble des sous-ensembles de  $\mathcal{V}$  — l'ensemble puissance — et  $\emptyset$  est l'ensemble vide. Comme pour les graphes,  $\psi : \mathcal{E} \rightarrow \mathbb{N}$  donne la multiplicité des hyperliens.

À l'instar des noeuds, il sera utile d'étiqueter tous les hyperliens du multiensemble  $\tilde{\mathcal{E}}$  par une valeur entière *unique*  $j \in \{1, 2, \dots, N'\}$ , où  $N'$  est le nombre total d'hyperliens. Par abus de notation, on écrira  $j \in \tilde{\mathcal{E}}$  pour désigner un hyperlien. Selon cette notation, chaque hyperlien  $j$  est associé à un ensemble  $e_j \in \mathcal{E}$ , mais chaque  $e_j$  n'est pas nécessairement unique s'il y a des hyperliens multiples.

Il est important de mentionner qu'il existe un isomorphisme entre les hypergraphes et un sous-ensemble des graphes *bipartis*, tel qu'illustré à la figure 1.2. Cet isomorphisme permet de recycler plusieurs des méthodes développées pour les graphes bipartis dans le contexte des hypergraphes. Cette relation est utilisée abondamment dans cette thèse et on favorise l'une ou l'autre de ces perspectives selon le contexte. Bien qu'un graphe biparti ne possède pas de boucles par définition<sup>5</sup>, les liens multiples sont possibles : on parlera alors d'un multigraphe biparti ou d'un graphe biparti simple si le graphe possède ou non des liens multiples et qu'il est nécessaire de préciser.

Dans sa forme la plus contrainte, un graphe biparti simple  $\mathcal{G} = (\mathcal{V}_1, \mathcal{V}_2, \mathcal{E})$  est défini par deux ensembles de noeuds disjoints,  $\mathcal{V}_1$  et  $\mathcal{V}_2$ , et un ensemble de liens  $\mathcal{E} \subseteq \{\{i, j\} | (i, j) \in \mathcal{V}_1 \times \mathcal{V}_2\}$ .

La transformation d'un hypergraphe multiple en un graphe biparti,  $\mathcal{H} \mapsto \mathcal{G}$ , mène toujours à un graphe biparti simple. Il faut d'abord associer l'ensemble des noeuds de l'hypergraphe au premier ensemble

4. Un complexe simplicial *abstrait* — on ignore ici l'aspect géométrique — est un type particulier d'hypergraphe pour lequel chaque sous-ensemble d'un hyperlien est aussi un hyperlien. Dans le contexte des réseaux aléatoires (voir section 1.3), la distinction entre un hypergraphe et un complexe simplicial n'a pas réellement d'importance dans la limite  $N \rightarrow \infty$ , car la probabilité que deux hyperliens d'un hypergraphe partagent plus d'un noeud devient négligeable. On peut alors interpréter un hypergraphe comme un complexe simplicial — en supposant que chaque sous-ensemble d'un hyperlien est aussi un hyperlien — et vice-versa.

5. Notons qu'une boucle d'un hypergraphe ne correspond pas à une boucle dans sa représentation bipartite.

de noeuds du graphe biparti,  $\mathcal{V} \mapsto \mathcal{V}_1$ . Ensuite, on associe un noeud  $j$  dans l'ensemble  $\mathcal{V}_2$  à chaque hyperlien  $j \in \tilde{\mathcal{E}}$  du multiensemble d'hyperliens, donc  $|\mathcal{V}_2| = N'$ . Finalement, on associe un lien  $\{i, j\}$  dans l'ensemble  $\mathcal{E}$  du graphe biparti pour chaque noeud  $i$  appartenant à l'ensemble  $e_j$  de l'hyperlien  $j$ . La transformation inverse est possible, mais uniquement pour les graphes bipartis simples pour lesquels chaque noeud de l'ensemble  $\mathcal{V}_2$  possède au moins un lien.

À noter que l'on favorise le terme *groupe* plutôt qu'hyperlien pour désigner les noeuds de l'ensemble  $\mathcal{V}_2$  lorsque l'on utilise la représentation en graphe biparti d'un hypergraphe. Comme à la figure 1.2, les groupes seront toujours illustrés par des noeuds noirs.

De manière analogue aux graphes, voici quelques concepts importants reliés aux hypergraphes.

**Composante** – La notion de composante se transpose aux hypergraphes : deux noeuds font partie de la même composante  $\mathcal{V}_\alpha \subseteq \mathcal{V}$  s'ils peuvent être rejoints par les hyperliens de l'hypergraphe. Similairement, les composantes séparées agissent comme des systèmes indépendants du point de vue dynamique.

**Voisinage** – On généralise le concept de voisinage en définissant  $\partial i$  comme l'ensemble des hyperliens  $j$  contenant le noeud  $i$ , c'est-à-dire  $\partial i = \{j \mid j \in \tilde{\mathcal{E}} \wedge i \in e_j\}$ . Du point de vue de la représentation en graphe biparti, cette définition du voisinage est équivalente à celle donnée précédemment pour les graphes, ici pour un graphe biparti simple —  $\partial i$  est l'ensemble des groupes  $j$  auxquels le noeud  $i$  est connecté.

**Matrice d'incidence ou de biadjacence** – On peut représenter la structure d'un hypergraphe à l'aide de la matrice d'incidence  $\mathbf{B}$  de dimension  $N \times N'$  où  $b_{ij} \in \{0, 1\}$  indique si le noeud  $i$  fait partie de l'hyperlien  $j$ . Cela fait en sorte que  $\mathbf{B}$  correspond aussi à la matrice de biadjacence du graphe biparti, où l'élément  $b_{ij} = 1$  indique que le noeud  $i$  appartient au groupe  $j$ . La somme de tous les éléments donne le nombre de liens du graphe biparti :

$$\sum_{i=1}^N \sum_{j=1}^{N'} b_{ij} = M . \quad (1.6)$$

**Hyperdegré ou appartenance** – L'hyperdegré  $m_i$  du noeud  $i$  correspond au nombre d'hyperliens qu'il possède, donc  $m_i = |\partial i|$ . On utilise aussi le terme *appartenance* dans le contexte de la représentation en graphe biparti pour désigner plus simplement le « degré d'appartenance » à des groupes. Comme le degré des noeuds dans un graphe, l'appartenance mesure à quel point chaque noeud est connecté au système. L'appartenance d'un noeud se calcule à partir de la matrice de biadjacence :

$$m_i = \sum_{j=1}^{N'} b_{ij} . \quad (1.7)$$

On définit la suite des appartenances  $(m_i)_{i=1}^N$ , ou sous forme vectorielle  $\mathbf{m} = [m_1, m_2, \dots, m_N]^\top$ . On note les appartenances minimale et maximale  $m_{\min}$  et  $m_{\max}$  respectivement.

**Taille d'un hyperlien ou d'un groupe** – La taille (ou cardinalité)  $n_j$  d'un hyperlien  $j$  est simplement la taille  $|e_j|$  de l'ensemble de noeuds  $e_j \in \mathcal{E}$ . De manière équivalente, la taille d'un groupe correspond au nombre de noeuds qui y est rattaché et se calcule à partir de la matrice de biadjacence :

$$n_j = \sum_{i=1}^N b_{ij} . \quad (1.8)$$

On définit la suite des tailles de groupe  $(n_j)_{j=1}^{N'}$ , ou sous forme vectorielle  $\mathbf{n} = [n_1, n_2, \dots, n_{N'}]^\top$ . On note les tailles de groupe minimale et maximale  $n_{\min}$  et  $n_{\max}$  respectivement.

**Distributions des appartenances et des tailles de groupe** – La distribution des appartenances  $P(m) \equiv g_m$  est la probabilité qu'un noeud sélectionné aléatoirement possède une appartenance  $m$ . Similairement, la distribution des tailles de groupe  $P(n) \equiv p_n$  est la probabilité qu'un groupe sélectionné aléatoirement soit de taille  $n$ . Les deux distributions respectent les conditions de normalisation

$$\sum_{m=m_{\min}}^{m_{\max}} g_m = \sum_{n=n_{\min}}^{n_{\max}} p_n = 1. \quad (1.9)$$

On définit également les moments d'ordre  $l$  comme suit :

$$\langle m^l \rangle = \sum_m m^l g_m, \quad \langle n^l \rangle = \sum_n n^l p_n, \quad (1.10)$$

où il est sous-entendu que l'on somme sur l'entièreté du support pour  $m$  et  $n$ . Ces deux distributions fondamentales,  $g_m$  et  $p_n$ , caractérisent la structure autant au niveau microscopique qu'au niveau mésoscopique, un aspect important pour l'étude des systèmes complexes.

**Coefficient d'agrégation biparti** – Le coefficient d'agrégation pour les graphes mesure à quel point les noeuds connectés partagent le même voisinage. Pour les hypergraphes, on veut mesurer à quel point deux noeuds appartenant à un même groupe appartiennent aux mêmes groupes en général. Il existe plusieurs mesurent permettant de généraliser cette notion [172, 277], mais la plus simple et naturelle que l'on utilise est celle donnée par Garry Robins et Malcolm Alexander dans le contexte des graphes bipartis [243] :

$$C_b = \frac{4N_{\boxtimes}}{N_{\mathbb{N}}}. \quad (1.11)$$

Le coefficient d'agrégation biparti  $C_b$  est le ratio entre le nombre de cycles de longueur quatre,  $N_{\boxtimes}$ , formant un motif de noeud papillon, et le nombre de parcours de longueur trois formant un motif en « N »,  $N_{\mathbb{N}}$ . Pour bien visualiser ces motifs, il est nécessaire de placer les noeuds des ensembles  $\mathcal{V}_1$  et  $\mathcal{V}_2$  sur deux lignes parallèles, comme à la figure 1.2. Le facteur quatre s'explique par le fait qu'il y a quatre parcours en « N » distincts dans un motif de noeud papillon.

### 1.2.1 Relation entre les hypergraphes et les graphes unipartis

Bien que la représentation en hypergraphe d'un système complexe soit plus générale, on pourrait vouloir la simplifier et projeter l'ensemble des interactions d'ordre supérieur vers des interactions dyadiques, c'est-à-dire transformer un hypergraphe en un graphe *uniparti*. L'approche typique est de transformer chaque hyperlien ou groupe de taille  $n$  en une *clique*, un sous-graphe complètement connecté. Le graphe résultant sera généralement un multigraphe : il suffit qu'il existe deux noeuds  $i$  et  $i'$  qui appartiennent à deux hyperliens  $j$  et  $j'$  pour créer des liens multiples.

Le chemin inverse n'est pas aussi simple, car la transformation  $\mathcal{H} \mapsto \mathcal{G}$  utilisant des cliques est une surjection. Différentes approches peuvent être utilisées pour transformer un graphe en un hypergraphe, par exemple en formulant l'objectif sous la forme d'un problème d'inférence et en utilisant des techniques d'échantillonnage [300].

### 1.3 Réseaux aléatoires

Pour modéliser un système complexe, il faut d’abord choisir une représentation mathématique adéquate. Ensuite, notre objectif est généralement de simplifier l’objet à l’étude afin de réaliser une analyse du système qui se trouve à être un compromis entre *fiabilité* et *interprétabilité*. Pour ce faire, on sélectionne un sous-ensemble des propriétés fondamentales du système afin de construire notre modèle, ce qui permet de faire abstraction des détails accessoires. Ces considérations sont valables autant pour l’élaboration du processus dynamique que pour la description de la structure même du système. C’est ici qu’entre en jeu la théorie des *réseaux aléatoires*.

On définit un modèle de réseaux aléatoires par une distribution de probabilité sur un ensemble de réseaux. Par exemple, on pourrait avoir la distribution  $P(\mathcal{G})$  sur l’ensemble de tous les graphes simples avec  $N$  noeuds. Pour définir  $P(\mathcal{G})$ , plusieurs approches existent selon l’objectif.

Une première approche, analogue à la mécanique statistique hors d’équilibre, est de voir le réseau comme un objet en évolution, c’est-à-dire un processus stochastique  $(\mathcal{G}(t))_{t \in \mathcal{D}_t}$ . La distribution  $P(\mathcal{G})$  est alors déterminée par les mécanismes du processus. C’est le cas notamment pour les modèles de croissance basés sur l’*attachement préférentiel* [16, 119, 121, 256]. Cette approche fut surtout utilisée initialement dans l’optique d’obtenir une meilleure compréhension des mécanismes ayant pu forger la structure des systèmes réels, mais récemment, les modèles de croissance furent exploités également dans le contexte de l’*archéologie* des réseaux [34, 42, 227, 302]. L’objectif est d’inférer les états passés du système uniquement à partir d’un réseau statique.

Une deuxième approche consiste à définir explicitement la probabilité de connexion entre chacun des noeuds du réseau. Par exemple, c’est le cas pour le modèle stochastique par blocs [4] et les graphes aléatoires de type Chung-Lu [57].

Une troisième approche, analogue dans ce cas-ci à la mécanique statistique à l’équilibre, s’appuie sur la théorie de l’information et le principe de la *maximisation de l’entropie* [145, 198]. L’idée est d’introduire une distribution de probabilité  $P(\mathcal{G})$  la plus objective possible, en injectant uniquement l’information nécessaire au respect d’un ensemble de contraintes. Lorsque ces contraintes doivent être respectées de manière exacte pour chaque réseau de l’ensemble, on parle alors d’un ensemble *microcanonique*, par opposition à un ensemble *canonique* lorsque les contraintes ne doivent être respectées qu’en moyenne.

Dans cette thèse, on se concentre sur les modèles de réseaux aléatoires basés sur un ensemble microcanonique, ceux typiquement utilisés pour l’étude des processus dynamiques. Pour les réseaux d’ordre supérieur, on utilise la représentation bipartite afin d’avoir une terminologie cohérente pour les graphes et les hypergraphes.

#### 1.3.1 Ensemble microcanonique avec demi-liens discernables

On définit une configuration  $\mathcal{C}$  d’un ensemble microcanonique comme un objet possédant la structure d’un réseau, mais chaque demi-lien est identifié par un indice unique. Pour un graphe,  $\mathcal{C} = (\mathcal{G}, \sigma)$ , où  $\sigma$  est une fonction qui étiquette chaque demi-lien du graphe  $\mathcal{G}$ . En d’autres termes, on considère chaque demi-lien comme *discernable* pour une configuration donnée. Par exemple, à la figure 1.3, les configurations  $\mathcal{C}_1$  et  $\mathcal{C}_2$  sont différentes, même si la structure du réseau résultant est identique (la matrice d’adjacence est la même). Cela implique qu’en général, plusieurs configurations sont associées



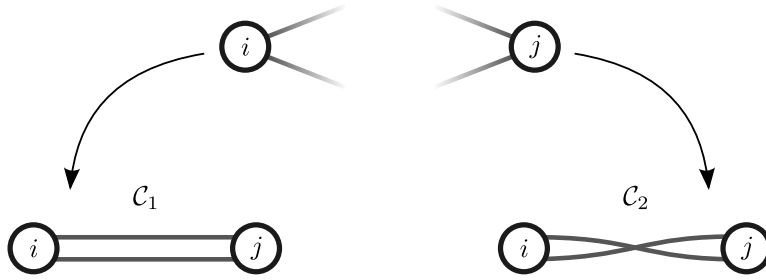


FIGURE 1.3 – Deux configurations différentes pour un multigraphe dues à la discernabilité des demi-liens.

à un même réseau.

Soit  $F$  un ensemble de configurations quelconque. En général, on limite l'ensemble des configurations possibles  $\Omega \subseteq F$  grâce à un certain nombre de contraintes à respecter. Soit  $\hat{\mathbf{y}} : F \rightarrow \mathbb{R}^q$  un vecteur de  $q$  mesures et  $\mathbf{y}$  le vecteur des valeurs à respecter. L'ensemble des configurations possibles correspond à  $\Omega = \{\mathcal{C} \mid \mathcal{C} \in F \wedge \hat{\mathbf{y}}(\mathcal{C}) = \mathbf{y}\}$ .

Finalement, la distribution de probabilité  $P(\mathcal{C})$  qui maximise l'entropie de Shannon est la distribution uniforme

$$P(\mathcal{C}) = \frac{1}{|\Omega|} . \quad (1.12)$$

Puisqu'un graphe  $\mathcal{G}$  est typiquement associé à plus d'une configuration,  $P(\mathcal{G})$  n'est pas nécessairement uniforme.

Choisir les demi-liens comme discernables permet de simplifier le processus de génération des réseaux ainsi que leurs propriétés analytiques. De plus, les propriétés statistiques d'un ensemble dont les demi-liens sont discernables ou indiscernables sont très semblables [30] et les différences entre les deux disparaissent généralement dans la limite  $N \rightarrow \infty$ . Lorsque l'indiscernabilité des demi-liens est essentielle, on suggère au lecteur les articles de Bailey K. Fosdick *et al.* [95] et de Philip S. Chodrow [55] pour les graphes et les hypergraphes respectivement.

On se concentre sur deux modèles de réseaux aléatoires : le modèle des configurations standard, générant des multigraphes avec boucles, et un modèle des configurations biparti permettant de générer des graphes bipartis simples.

### 1.3.2 Modèle des configurations standard

L'ensemble  $F$  pour le modèle des configurations standard est construit à partir de l'ensemble des multigraphes avec boucles de  $N$  noeuds — pour chaque graphe, il existe un certain nombre de configurations correspondant aux permutations des indices des demi-liens de chaque noeud. En fait, pour chaque graphe simple, le nombre de configurations possibles est simplement  $\prod_{i=1}^N k_i!$ ; pour un multigraphe avec boucles, l'expression est un peu plus compliquée [220]. On s'intéresse plus particulièrement aux graphes associés à une suite des degrés  $\mathbf{k}$  fixée, donc l'ensemble des configurations possibles correspond à  $\Omega = \{\mathcal{C} \mid \mathcal{C} \in F \wedge \hat{\mathbf{k}}(\mathcal{C}) = \mathbf{k}\}$ , avec  $\hat{\mathbf{k}}(\mathcal{C})$  correspondant à la suite des degrés de la configuration  $\mathcal{C}$ . La suite des degrés est totalement arbitraire, mais doit respecter la contrainte  $\sum_{i=1}^N k_i = 2M$ , ce qui implique que la somme des degrés est nécessairement paire.

Dans certains cas, on ne s'intéresse pas à une suite des degrés particulière, mais plutôt aux suites de longueur  $N$  tirées d'une distribution des degrés  $p_k$ . Dans ce cas, il suffit de tirer une suite  $\mathbf{k}$  de la distribution  $p_k$ . Si la somme des degrés est un nombre impair, on tire à nouveau le degré d'un des noeuds choisi aléatoirement jusqu'à ce que la somme soit un nombre pair. Le modèle de réseau aléatoire résultant n'est plus tout à fait un ensemble microcanonique, mais ce n'est pas non plus un ensemble canonique pour lequel le degré de chaque noeud serait plutôt fixé en moyenne [58].

La manière la plus simple de générer une configuration de l'ensemble  $\Omega$  est par échantillonnage directe [220] : on assigne  $k_i$  demi-lien à chaque noeud  $i \in \mathcal{V}$ , puis on sélectionne une paire de demi-liens uniformément parmi les  $2M$  demi-liens disponibles pour créer un lien. On répète ce processus jusqu'à l'épuisement des demi-liens, menant à une configuration  $\mathcal{C}$  et le graphe  $\mathcal{G}$  associé. On appelle cette technique le *jumelage des demi-liens*.

Le jumelage des demi-liens est exact et efficace pour échantillonner les multigraphes avec boucles, mais n'est pas approprié dans le contexte des graphes simples par exemple. Pour une revue exhaustive des méthodes d'échantillonnage pour les différentes versions du modèle des configurations, voir l'article de Fosdick *et al.* [95].

Une propriété fondamentale du modèle des configurations est que les degrés ne sont pas corrélés dans la limite  $N \rightarrow \infty$  [220]. La probabilité  $P(k, k')$  d'avoir un lien joignant un noeud de degré  $k$  et un autre de degré  $k'$  est

$$P(k, k') = \left( \frac{k p_k}{\langle k \rangle} \right) \left( \frac{N k' p_{k'} - \delta_{k, k'}}{N \langle k \rangle - 1} \right) \simeq \frac{k k' p_k p_{k'}}{\langle k \rangle^2}. \quad (1.13)$$

Par conséquent, la probabilité  $P(k'|k)$  d'arriver à un noeud de degré  $k'$  si on suit un lien aléatoire partant d'un noeud de degré  $k$  est

$$P(k'|k) = \frac{P(k, k')}{\sum_{k'} P(k, k')} \simeq \frac{k' p_{k'}}{\langle k \rangle}, \quad (1.14)$$

et donc, ne dépend pas du degré du noeud initial.

### 1.3.3 Modèle des configurations biparti

L'ensemble  $F$  pour le modèle des configurations biparti est construit à partir de l'ensemble des graphes bipartis simples de  $N$  noeuds et  $N'$  groupes, tous de tailles  $n \geq 1$ . Pour chaque graphe biparti simple, il y a  $\prod_{i=1}^N m_i! \prod_{j=1}^{N'} n_j!$  configurations possibles. Grâce à l'isomorphisme discuté à la section 1.2, cela correspond aussi à l'ensemble des hypergraphes avec boucles et hyperliens multiples. On s'intéresse plus particulièrement aux graphes bipartis ayant une suite des appartenances  $\mathbf{m}$  et une suite des tailles de groupe  $\mathbf{n}$  fixées. L'ensemble des configurations possibles correspond alors à

$$\Omega = \{ \mathcal{C} | \mathcal{C} \in F \wedge \hat{\mathbf{m}}(\mathcal{C}) = \mathbf{m} \wedge \hat{\mathbf{n}}(\mathcal{C}) = \mathbf{n} \}, \quad (1.15)$$

avec  $\hat{\mathbf{m}}(\mathcal{C})$  et  $\hat{\mathbf{n}}(\mathcal{C})$  correspondant respectivement à la suite des appartenances et à la suite des tailles de groupe de la configuration  $\mathcal{C}$ .

Pour générer un graphe biparti à partir des suites  $\mathbf{m}$  et  $\mathbf{n}$ , il faut d'abord que leur somme soit la même — correspondant à  $M$  — pour avoir autant de demi-liens associés aux noeuds qu'aux groupes. Par contre, une contrainte supplémentaire est nécessaire pour garantir que les suites puissent générer un

graphe biparti *simple*. Le théorème de Gale-Ryser [96, 246] suivant donne les conditions nécessaires et suffisantes à la réalisation d'un graphe biparti simple à partir d'une suite des appartenances et une suite des tailles de groupe données.

**Théorème 1.3.1** *Soit la suite des appartenances  $(m_i)_{i=1}^{\tilde{N}}$  ordonnée en ordre décroissant ( $m_1 \geq m_2 \geq \dots \geq m_{\tilde{N}}$ ) et la suite des tailles de groupe  $(n_j)_{j=1}^{\tilde{N}}$ , respectant les conditions suivantes :*

- (i)  $\sum_{i=1}^{\tilde{N}} m_i = \sum_{j=1}^{\tilde{N}} n_j$  ;
- (ii)  $\sum_{i=1}^d m_i \leq \sum_{j=1}^{\tilde{N}} \min(n_j, d)$  pour tout  $d \in \{1, \dots, \tilde{N}\}$ .

*Par conséquent, la paire de suites  $(m_i)_{i=1}^{\tilde{N}}$  et  $(n_j)_{j=1}^{\tilde{N}}$  est bigraphique : il est possible de construire un graphe biparti simple ayant cette suite des appartenances et cette suite des tailles de groupe. Réciproquement, tout graphe biparti simple doit respecter les conditions précédentes.*

En général, le nombre de noeuds  $N$  est différent du nombre de groupes  $N'$ , mais il suffit de prendre  $\tilde{N} = \max(N, N')$  et de rajouter des zéros à la fin de la plus courte des deux suites afin qu'elle soit de longueur  $\tilde{N}$ .

Encore une fois, on s'intéresse parfois à des distributions des appartenances  $g_m$  et des tailles de groupe  $p_n$  plutôt qu'à des suites  $\mathbf{m}$  et  $\mathbf{n}$  particulières. Dans ce cas, on fait le choix de fixer le nombre de noeuds  $N$  et de laisser fluctuer le nombre de groupes  $N'$ . On tire une suite des appartenances à partir de la distribution  $g_m$ , ce qui fixe aussi le nombre de liens  $M$ . Ensuite, on tire une suite  $\mathbf{n}$ . On ajoute ou on enlève un groupe de manière aléatoire jusqu'à ce que les deux conditions du théorème 1.3.1 soient satisfaites.

Afin de générer un graphe biparti de telle sorte que la configuration  $\mathcal{C}$  soit tirée uniformément de l'ensemble  $\Omega$ , on utilise l'échantillonnage par chaîne de Markov. L'idée est de créer un processus stochastique  $(\mathcal{C}(t))_{t=1}^T$  dont la distribution stationnaire de la chaîne, notée  $P^*(\mathcal{C})$ , correspond à la distribution uniforme sur l'ensemble  $\Omega$ .

D'abord, il faut définir la probabilité de transition  $P(\mathcal{C}'|\mathcal{C})$ . Supposons que l'on choisisse aléatoirement deux liens  $e_1 = (i_1, j_1)$  et  $e_2 = (i_2, j_2)$  dans l'ensemble  $\mathcal{E}$  et que l'on échange les groupes associés aux deux liens,  $j_1 \leftrightarrow j_2$ , créant ainsi deux nouveaux liens  $e'_1 = (i_1, j_2)$  et  $e'_2 = (i_2, j_1)$ . Si ce changement ne crée pas de lien multiple, on accepte la transformation, sinon on conserve les liens  $e_1$  et  $e_2$ . On appelle ce type de transformation un *échange de liens* [95], et il fut montré que n'importe quelle paire de configurations de l'ensemble  $\Omega$  peut être reliée par un certain nombre d'échanges de liens — en termes plus formels la chaîne de Markov est *irréductible* [82, 246]. De plus, cette chaîne de Markov est *apériodique*, car l'échange de liens n'est pas toujours accepté afin d'éviter les liens multiples, auquel cas  $\mathcal{C}' = \mathcal{C}$ . Ces deux propriétés font en sorte que la chaîne de Markov est *ergodique*, et donc qu'il est possible d'échantillonner la chaîne de Markov dans la limite  $T \rightarrow \infty$  pour approximer la distribution stationnaire.

Si une configuration  $\mathcal{C}$  peut être transformée en une configuration  $\mathcal{C}'$  à l'aide d'un seul échange de lien, alors la probabilité de transition est  $P(\mathcal{C}'|\mathcal{C}) = 1/\binom{M}{2}$ . La probabilité de transition inverse  $P(\mathcal{C}|\mathcal{C}')$  étant identique, cela implique que la chaîne de Markov est *réversible* par le critère de Kolmogorov. Par

conséquent, la distribution stationnaire respecte la condition de *bilan détaillé*

$$P^*(\mathcal{C})P(\mathcal{C}'|\mathcal{C}) = P^*(\mathcal{C}')P(\mathcal{C}|\mathcal{C}') , \quad (1.16a)$$

$$\Rightarrow P^*(\mathcal{C}) = P^*(\mathcal{C}') . \quad (1.16b)$$

Puisque la chaîne est irréductible, la distribution stationnaire  $P^*(\mathcal{C})$  est alors uniforme.

L'approche par chaîne de Markov est plus flexible que le jumelage des demi-liens pour échantillonner des réseaux aléatoires [55, 95], mais il faut maintenant composer avec des défis supplémentaires afin de générer un échantillon de configurations indépendamment distribuées de  $P^*(\mathcal{C})$  :

1. il faut choisir une configuration initiale  $\mathcal{C}(0)$  pour partir la chaîne de Markov ;
2. il faut laisser la chaîne évoluer suffisamment longtemps pour approximer l'état stationnaire ;
3. entre chaque mesure, il faut encore une fois laisser la chaîne évoluer suffisamment longtemps pour que les mesures soient indépendantes.

Pour le point 1, on peut soit utiliser un graphe connu ayant les suites  $\mathbf{m}$  et  $\mathbf{n}$  désirées ou utiliser la technique de jumelage des demi-liens, créant ainsi un multigraphe biparti. Il suffit alors d'effectuer des échanges de liens ciblés jusqu'à l'élimination des liens multiples. Les points 2 et 3 sont quant à eux reliés au concept de *temps de mélange* discuté plus bas.

Finalement, on présente certaines propriétés analytiques du modèle des configurations biparti. Si les liens multiples étaient permis, la probabilité qu'un lien soit relié à un noeud d'appartenance  $m$  et à un groupe de taille  $n$  serait simplement

$$P(m, n) = \frac{mng_m p_n}{\langle m \rangle \langle n \rangle} . \quad (1.17)$$

Par contre, cela ne tient pas compte de la contrainte d'avoir un graphe biparti simple, laquelle introduit des corrélations. Ces corrélations sont toutefois négligeables si le nombre moyen de liens entre chaque noeud et chaque groupe tend vers 0 dans la limite  $M \rightarrow \infty$ . Cela est respecté si

$$\frac{m_{\max} n_{\max}}{M} \ll 1 \iff m_{\max} n_{\max} = o(M) . \quad (1.18)$$

Sous ces conditions, pour les graphes de l'ensemble  $\Omega$ ,

$$P(m, n) \simeq \frac{mng_m p_n}{\langle m \rangle \langle n \rangle} . \quad (1.19)$$

Par conséquent, la probabilité d'arriver à un noeud de degré  $m$  si on prend un noeud aléatoire dans un groupe de taille  $n$  est

$$P(m|n) = \frac{P(m, n)}{\sum_m P(m, n)} \simeq \frac{mg_m}{\langle m \rangle} . \quad (1.20)$$

Similairement, la probabilité d'arriver à un groupe de taille  $n$  si on choisit aléatoirement un groupe auquel appartient un noeud d'appartenance  $m$  est

$$P(n|m) = \frac{P(m, n)}{\sum_n P(m, n)} \simeq \frac{np_n}{\langle n \rangle} . \quad (1.21)$$

Il n'y a donc pas de corrélations entre l'appartenance d'un noeud et la taille des groupes auxquels il appartient.

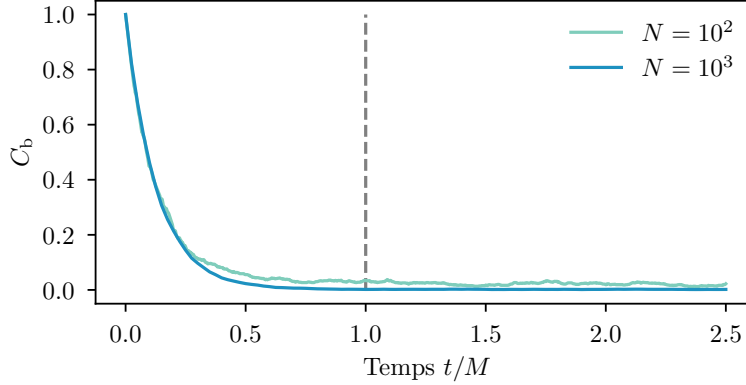


FIGURE 1.4 – Temps de mélange pour la chaîne de Markov associée à l’échantillonnage du modèle des configurations biparti. On utilise  $g_m = \delta_{m,4}$  et  $p_n = \delta_{n,4}$  et la configuration initiale  $\mathcal{C}(0)$  est obtenue en créant uniquement des bicliques de quatre noeuds et de quatre groupes. On s’intéresse à l’évolution du coefficient d’agrégation biparti  $C_b$  en fonction du temps, soit le nombre de tentatives d’échange de liens. La ligne pointillée verticale indique lorsque  $t = M$ .

### 1.3.4 Temps de mélange

En échantillonnant les réseaux par une chaîne de Markov, le nombre de tentatives d’échange de liens correspond au temps  $t$  de la chaîne. On définit le *temps de mélange* ou de décorrélation  $T$  comme étant le temps nécessaire pour que la distribution de la chaîne de Markov  $P[\mathcal{C}(T)]$  soit suffisamment près de la distribution stationnaire  $P^*(\mathcal{C})$ , indépendamment de la configuration initiale. On pourrait définir plus formellement le temps de mélange par un critère précis, mais cette définition intuitive est suffisante pour fixer  $T$  de manière approximative.

Une chaîne de Markov se mélange rapidement si  $T$  est borné supérieurement par un polynôme en  $N$ . Il existe certains résultats théoriques qui permettent de caractériser le temps de mélange des chaînes de Markov pour l’échantillonnage de réseaux aléatoires, mais la plupart du temps ces résultats sont limités à des cas particuliers. Par exemple, pour le cas des graphes bipartis simples, si  $g_m = \delta_{m,m'}$  ou  $p_n = \delta_{n,n'}$ , alors la chaîne de Markov que l’on a décrite précédemment se mélange rapidement [82]. À notre connaissance, il n’existe pas de résultats théoriques pour des distributions  $g_m$  et  $p_n$  arbitraires.

En pratique, on caractérise le temps de mélange en évaluant le temps nécessaire à la convergence d’un certain nombre de mesures scalaires  $\Psi : \mathcal{C} \rightarrow \mathbb{R}$  [99]. À la figure 1.4, on montre la convergence du coefficient d’agrégation biparti pour des réseaux de taille différente. Au temps  $t = 0$ , on choisit une configuration atypique de l’ensemble maximisant le coefficient d’agrégation biparti. Pour ce faire, on connecte les noeuds et les groupes en *bicliques*<sup>6</sup> de quatre noeuds et de quatre groupes. Malgré cela, on remarque que le coefficient d’agrégation converge en un temps  $T = O(M)$ .

Pour des distributions  $g_m$  et  $p_n$  hétérogènes, il est possible que le temps de mélange ne soit pas  $O(M)$ . Toutefois, il est possible d’accélérer la convergence en utilisant des configurations initiales  $\mathcal{C}(0)$  tirées d’une distribution qui se rapproche de la distribution uniforme, par exemple en utilisant le jumelage de demi-liens et en éliminant les liens multiples par des échanges de liens. Bien que la distribution  $P(\mathcal{C})$  associée au jumelage de demi-liens ne soit pas uniforme, si le nombre de liens multiples n’est pas

6. Sous-graphes bipartis maximalelement connectés ou chaque noeud est connecté à tous les groupes et vice-versa.

trop grand, on peut se convaincre que cette distribution est assez près de la distribution uniforme.

Une fois que la distribution de la chaîne s'est suffisamment rapprochée de la distribution stationnaire, plusieurs techniques d'échantillonnage sont possibles. On peut mesurer une configuration à chaque temps  $t$ , ou mesurer une configuration de manière périodique après un temps de décorrélation  $T$ , ce qui permet de considérer chaque configuration mesurée comme une réalisation indépendante tirée de  $P^*(\mathcal{C})$ .

En pratique, on utilise l'approche suivante pour générer un échantillon de  $q$  configurations indépendantes tirées de  $P^*(\mathcal{C})$  : on crée  $q$  chaînes de Markov à partir de configurations  $\mathcal{C}(0)$  différentes obtenues par jumelage de demi-liens. Ensuite, on laisse évoluer chaque chaîne un temps  $T = O(M)$  avant d'échantillonner la configuration finale  $\mathcal{C}(T)$ .

## 1.4 Reconstruction et propriétés statistiques des systèmes sociaux

Il existe plusieurs approches pour *reconstruire* un réseau social. On suppose ici qu'il existe un réseau social réel sous-jacent, et à partir de mesures — la majorité du temps indirectes — on crée un réseau que l'on appelle une *reconstruction*. Récemment, il y eut plusieurs avancées théoriques sur le sujet, notamment en faisant usage de l'inférence statistique [219, 236, 298] pour reconstruire les réseaux à partir de données bruitées et quantifier l'incertitude sur l'existence des liens.

La reconstruction d'un réseau social nécessite d'abord l'accumulation de données, le plus souvent en questionnant directement les individus concernés [220]. Pour ce faire, on peut questionner l'ensemble de la population d'intérêt si c'est raisonnable et possible de le faire. Étant donné l'effort requis par une telle tâche, cette approche pour recueillir des données est souvent limitée à de petites communautés [240, 241].

Il est aussi possible d'utiliser une méthode d'échantillonnage pour explorer le tissu social, souvent en exploitant le réseau lui-même pour découvrir de nouveaux individus potentiels — ce qui est notamment utile pour recueillir de l'information sur les populations « cachées », stigmatisées ou réticentes à se faire questionner, comme les utilisateurs de drogues injectables [212, 245]. Entre autres, l'échantillonnage en boule de neige est une technique standard : on questionne un individu sur, par exemple, ses relations d'amitié ou avec qui il s'injecte de la drogue. Les réponses servent à cumuler de nouveaux candidats à questionner et ainsi de suite.

La recherche des contacts en épidémiologie se fait de manière analogue à l'échantillonnage en boule de neige. À partir d'un individu infecté, on tente de retrouver les autres individus ayant pu avoir des contacts avec cette personne infectée. On peut faire du traçage *en amont* pour essayer de retrouver la source de l'infection, ou du traçage *en aval* afin de prévenir la propagation. Les données recueillies mènent naturellement à un un graphe en arbre orienté indiquant qui a infecté qui. Selon le mode de propagation, la recherche de contact peut aussi se faire en recueillant l'information sur les *lieux* visités par l'individu infecté — c'est notamment une technique employée en ce moment pour combattre la pandémie de COVID-19 [3], puisque la maladie est aéroportée. Dans ce cas, les données de transmission prennent la forme d'un graphe biparti. Par contre, il est important de mentionner que les réseaux alors reconstruits sont des échantillons extrêmement biaisés du tissu social réel sur lequel évolue le processus de contagion. La recherche des contacts est alors surtout utilisée pour estimer la *distribution des cas*

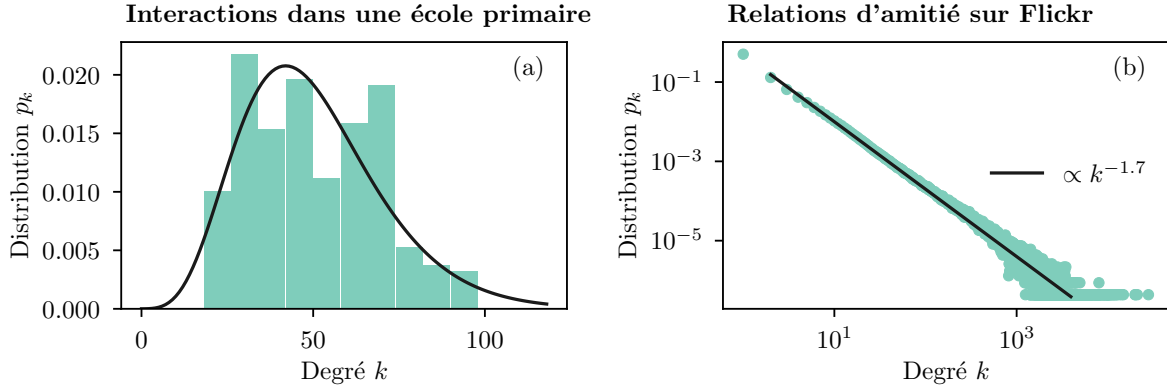


FIGURE 1.5 – Distribution des degrés pour des graphes extraits de jeux de données. (a) Interactions entre les étudiants et les enseignants d’une école primaire de Lyon en France durant une journée complète [271]. Les interactions sont identifiées à l’aide d’un système de radio-identification (*RFID*). La distribution des degrés est extraite du graphe obtenu en mettant un lien entre chaque paire d’individus ayant eu au moins une interaction. La ligne noire correspond à une distribution binomiale négative ayant la même moyenne et la même variance que les données,  $\langle k \rangle \approx 50$  et  $\text{Var}(k) \approx 358$ , ce qui lui vaut un indice de dispersion de  $D \approx 7.16$ . (b) Relations d’amitié déclarées entre les utilisateurs du site web Flickr pour le partage de photos et de vidéos selon des données accumulées en 2006 et en 2007 [204]. La distribution des degrés est extraite du graphe obtenu en transformant chaque lien orienté en lien non orienté. Le comportement de la distribution pour les hauts degrés s’apparente à une loi de puissance de la forme  $p_k \sim k^{-\gamma}$  avec  $\gamma \approx 1.7$ .

*secondaires* — la distribution du nombre de nouvelles infections causées par un seul individu infecté — et non le réseau de contacts lui-même.

Avec l’avènement de la collecte massive des données, le traçage des téléphones mobiles et les autres apports de l’ère numérique, il est possible d’estimer le nombre de contacts potentiels entre individus comme jamais auparavant, le tout sans avoir à questionner les individus concernés. On peut, par exemple, mesurer directement les interactions sociales entre individus en utilisant des capteurs [48, 271], identifier les liens d’amitié ou l’appartenance à des groupes grâce aux médias sociaux [204, 205], ou utiliser la géolocalisation des téléphones mobiles pour déterminer la fréquentation de certains endroits comme les restaurants ou les bars [296]. Ces données se prêtent naturellement à une représentation réseau, soit sous la forme d’un graphe standard ou biparti, et incluent généralement des métadonnées riches sur les utilisateurs, ce qui apporte une compréhension en profondeur du système social. Bien entendu, ce type de collecte de données soulève également son lot d’enjeux éthiques [306].

Malgré l’imperfection des données et de la reconstruction des réseaux sociaux, on peut regrouper, caractériser et comparer les réseaux à partir de leurs propriétés statistiques. On présente ici sommairement quelques propriétés *universelles*<sup>7</sup> ayant une incidence majeure sur les processus dynamiques.

### 1.4.1 Structure hétérogène

Pour les graphes, le degré  $k_i$  d’un noeud  $i$  détermine en grande partie l’état dynamique de ce dernier. Par exemple, si on considère un processus de contagion, plus un individu est connecté à la population (plus son degré est élevé) plus il est probable qu’il se fasse éventuellement infecter. Par conséquent,

7. Partagées par une portion considérable des systèmes.

la distribution des degrés  $p_k$  a un effet considérable sur l'état global du système. Bien entendu, le degré moyen  $\langle k \rangle$  est important, mais les propriétés critiques d'un système dépendent souvent des deux premiers moments de la distribution des degrés [19].

Par conséquent, on s'intéresse généralement à l'*hétérogénéité* des degrés, que l'on quantifie notamment grâce à l'indice de dispersion  $D = \text{Var}(k)/\langle k \rangle$ , où  $\text{Var}(k) = \langle k^2 \rangle - \langle k \rangle^2$  est la variance de la distribution  $p_k$ . Plus l'indice de dispersion est élevé, plus les degrés sont hétérogènes. Par exemple, selon cette mesure, une distribution des degrés *régulière*  $p_k = \delta_{k,k_0}$  aura un indice de dispersion nul. Pour un réseau aléatoire de type Erdős-Rényi dont seul le nombre de liens total est fixé (ou la probabilité d'un lien), la distribution des degrés est asymptotiquement une loi de Poisson dans la limite  $N \rightarrow \infty$ ,

$$p_k \simeq \frac{\langle k \rangle^k e^{-\langle k \rangle}}{k!} \quad \forall k \geq 0. \quad (1.22)$$

Dans ce cas, l'indice de dispersion est de 1. Puisqu'en s'éloignant d'une distribution de Poisson cela indique une certaine forme d'organisation non triviale, on qualifie d'*homogène* les distributions ayant un indice de dispersion  $D \leq 1$  et d'*hétérogène* toutes celles possédant un indice de dispersion  $D > 1$ .

Il existe plusieurs distributions standards qui permettent de modéliser un comportement hétérogène, comme la loi binomiale négative :

$$p_k = \frac{\Gamma(k+r)}{\Gamma(k)\Gamma(r+1)} q^k (1-q)^r. \quad (1.23)$$

Dans la limite  $r \rightarrow \infty$  on retombe sur la loi de Poisson, tandis qu'une valeur  $r < \infty$  mène à une distribution plus hétérogène. La loi binomiale négative est couramment utilisée en épidémiologie pour représenter la distribution pour le nombre de cas secondaires [185], et est fort utile dans une multitude de situations pour lesquelles la distribution possède un mode bien défini, mais dont l'indice de dispersion est supérieur à 1. C'est le cas à la figure 1.5(a) pour la distribution des degrés associée aux interactions entre les individus d'une école primaire.

Toutefois, c'est l'universalité des distributions des degrés avec un comportement asymptotique en loi de puissance qui a attiré l'attention en science des réseaux, soit

$$p_k \sim k^{-\gamma}, \quad (1.24)$$

pour  $k \rightarrow \infty$ . On dit de ces graphes qu'ils sont *sans échelle*, car leur distribution des degrés n'a pas d'échelle caractéristique<sup>8</sup>.

L'universalité de ces distributions est souvent expliquée à partir du mécanisme d'attachement préférentiel pour la croissance des systèmes complexes [16, 121, 256], lequel favoriserait l'émergence des distributions en loi de puissance. Malgré un désaccord récent sur la définition des réseaux sans échelle [17, 39, 282], leur omniprésence est un fait généralement accepté par la communauté de la science des réseaux. Selon une étude récente [282], la distribution des degrés de près de 50% des graphes non orientés<sup>9</sup> pourrait avoir un comportement asymptotique en loi de puissance. On présente à la figure 1.5(b) une distribution des degrés typique d'un graphe sans échelle, dans ce cas-ci relié au nombre d'« amis » que possèdent les utilisateurs du site web Flickr.

8. L'analogie provient de la physique statistique. Le comportement des paramètres d'ordre près des points critiques est décrit par des fonctions homogènes, qui, par définition, n'ont pas d'échelle caractéristique. Par exemple, si on a une densité de probabilité  $\varphi(x) \propto x^\alpha$ , un changement de variable  $y = ax$  mène à une densité de la même forme.

9. Parmi les graphes accessibles, reconstruits à partir de jeux de données et sélectionnés dans leur étude.



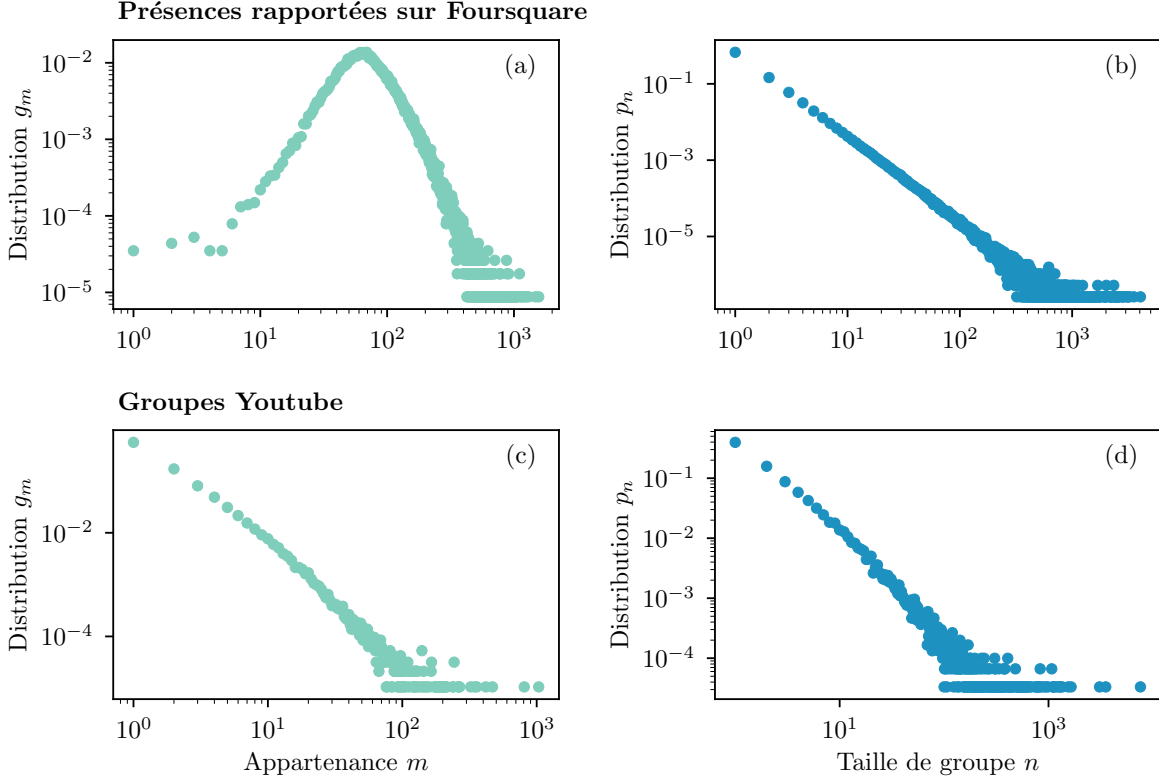


FIGURE 1.6 – Distribution des appartenances et des tailles de groupes pour des graphes bipartis simples extraits de jeux de données. (a)-(b) Présences dans des lieux (restaurant, bar, événement, etc.) rapportées par des utilisateurs de l’application mobile Foursquare [296]. On associe un groupe à chaque lieu et chaque utilisateur ayant rapporté sa présence dans ce lieu appartient au groupe. L’appartenance d’un utilisateur est le nombre de présences dans des lieux différents et la taille d’un groupe est le nombre de personnes associées à ce lieu. (c)-(d) Appartenance à des groupes sur la média social YouTube [205].

Une des conséquences de la distribution des degrés en loi de puissance est que le système contient une fraction non négligeable de noeuds de très haut degré. Ces noeuds ont un rôle extrêmement important sur la dynamique des systèmes complexes — en épidémiologie, ce sont les individus que l’on associe aux événements de *superdissémination*, lors desquels un grand nombre de nouveaux cas est causé par une seule personne infectée. Un autre aspect important des distributions en loi de puissance est que certains des moments de  $p_k$  ne sont pas définis dans la limite  $k_{\max} \rightarrow \infty$ . Par exemple, pour  $2 < \gamma < 3$ , la moyenne  $\langle k \rangle$  est définie, mais le second moment  $\langle k^2 \rangle$  diverge (ce qui implique  $D \rightarrow \infty$  également). Cela a des conséquences majeures, notamment sur les propriétés critiques des processus dynamiques [19].

Pour les systèmes complexes mieux représentés par des interactions d’ordre supérieur, les jeux de données disponibles sont souvent sous la forme d’un graphe biparti, et donc on favorise cette perspective.

À l’instar de la distribution des degrés pour les graphes, les distributions des appartenances  $g_m$  et des tailles de groupe  $p_n$  ont une influence importante sur l’état dynamique des systèmes. De plus, l’universalité des distributions en loi de puissance s’applique également au niveau mésoscopique. Les données empiriques disponibles suggèrent que l’appartenance des noeuds et la taille des groupes sont

souvent très hétérogènes, avec

$$g_m \sim m^{-\gamma_m} \quad \text{et} \quad p_n \sim n^{-\gamma_n}, \quad (1.25)$$

pour  $m \rightarrow \infty$  et  $n \rightarrow \infty$ . Encore une fois, le mécanisme d’attachement préférentiel permettrait d’expliquer l’omniprésence de ces structures hétérogènes [119].

Plus spécifiquement, pour 35% des graphes bipartis, les deux distributions  $g_m$  et  $p_n$  auraient un comportement asymptotique en loi de puissance, et dans 74% des cas, au moins l’une des deux distributions serait asymptotiquement en loi de puissance [282]. À la figure 1.6, on présente deux exemples de systèmes avec des distributions  $g_m$  et  $p_n$  asymptotiquement en loi de puissance. Le premier concerne la mobilité des individus se rapportant dans des lieux sur l’application Foursquare, tandis que le deuxième concerne les interactions sociales en ligne des utilisateurs de YouTube.

Aux chapitres 3, 4 et 6 de cette thèse, nous étudions les conséquences d’avoir des distributions  $g_m$  et  $p_n$  asymptotiquement en loi de puissance sur les processus de contagion.

#### 1.4.2 Contacts en rafales et durée des contacts hétérogène

Au-delà de la structure des interactions, la dynamique des interactions elles-mêmes affecte de manière significative l’évolution d’un système complexe, une perspective souvent étudiée à l’aide des réseaux temporels [134, 135]. Bien que dans cet ouvrage on se limite aux réseaux statiques, il est possible d’intégrer dans nos modèles certaines des propriétés temporelles des interactions entre individus connectés. Par exemple, les liens ou les hyperliens d’un réseau peuvent représenter les relations entre les individus d’une population, mais le moment exact et la durée des *contacts* entre individus peuvent varier énormément.

On définit plus formellement un contact par un couple  $(t, \tau)$ , où  $t$  est le moment auquel ce contact est initié et  $\tau$  est la durée du contact. La signification physique d’un contact n’est pas unique : ce peut être la rencontre physique d’un nombre d’individus, une discussion, un événement, etc. Pour les processus de contagion, on suppose que l’agent pathogène ou l’information est transmis pendant ces contacts entre individus.

On s’intéresse généralement à deux densités de probabilité,  $\varphi(\tau)$  pour la durée des contacts et  $\varphi(\Delta t)$  pour le temps  $\Delta t$  entre deux contacts. Comme illustré à la figure 1.7, la durée des contacts et le temps entre les contacts sont parfois très hétérogènes, avec un comportement asymptotique en loi de puissance. Ceci est très différent de ce que l’on pourrait s’attendre si on supposait par exemple un processus de Poisson, lors duquel un contact est initié à un taux constant  $\alpha$  et, une fois initié, le contact se termine à un taux  $\beta$ . Dans ce cas, on aurait des densités de probabilité de forme exponentielle  $\varphi(\tau) \propto e^{-\beta\tau}$  et  $\varphi(\Delta t) \propto e^{-\alpha\Delta t}$ , donc beaucoup plus homogènes. L’émergence de ces distributions peut s’expliquer, entre autres, par la *mémoire* des individus du système [152, 153, 270], auquel cas le processus stochastique décrivant les contacts est non markovien.

Un temps entre deux contacts  $\Delta t$  hétérogène implique des contacts en *rafales*, c’est-à-dire des successions de contacts rapprochés temporellement séparés par de longues périodes d’inactivité [152]. Leur effet sur les processus de contagion fut étudié dans plusieurs articles [148, 154, 280]. Au chapitre 5, on s’intéresse plutôt à l’influence de l’hétérogénéité de la durée des contacts  $\tau$  dans le contexte des contagions sur réseaux d’ordre supérieur.

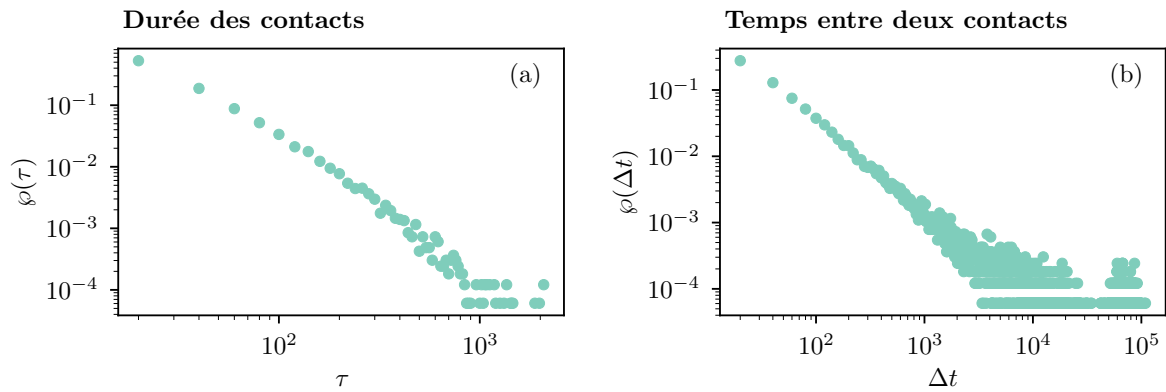


FIGURE 1.7 – Propriétés statistiques des contacts entre individus durant une conférence à Nice (France) en 2009 [100]. Les contacts sont mesurés à l’aide d’un système de radio-identification (*RFID*). Le temps est séparé en intervalles de 20 secondes et un contact est enregistré entre deux individus s’ils étaient face-à-face à une distance de moins de 1.5 mètre. (a) Durée des contacts : si un contact est enregistré sur plusieurs intervalles de temps consécutifs, on considère que c’est un seul contact d’une durée prolongée. (b) Temps entre deux contacts : pour chaque paire d’individus, on mesure le nombre d’intervalles consécutifs sans contact entre les deux individus.

## Chapitre 2

# Processus de contagion

Plusieurs considèrent que les premiers modèles mathématiques en épidémiologie furent développés par Daniel Bernoulli, publiés dans un ouvrage en 1766 [28] dans lequel il traite des avantages de la variolisation, un ancêtre de la vaccination. Depuis, plusieurs articles fondateurs ont contribué à établir la modélisation des processus de contagion comme un champ de recherche actif en mathématiques appliquées. Entre autres, mentionnons les travaux de William O. Kermack et Anderson G. McKendrick [72, 160–162], les premiers en 1927, sur l'étude des modèles de type Susceptible-Infecté-Retiré (SIR), lesquels servent toujours de fondements pour la plupart des modèles en épidémiologie à ce jour. Il y a aussi Thomas C. Schelling [251] en 1971 et Mark Granovetter [112] en 1978 qui introduisent les dynamiques de seuil, à la base de la plupart des modèles de contagion sociale. Le livre de Roy M. Anderson et Robert May [12], publié en 1991, demeure une référence incontournable pour la modélisation des maladies infectieuses. Finalement, mentionnons l'article de Romualdo Pastor-Satorras et Alessandro Vespignani [233] en 2001, lequel fut un important pas en avant pour établir le domaine de l'épidémiologie sur réseaux en mettant de l'avant l'important rôle de l'hétérogénéité des interactions sociales.

On aborde ce chapitre en précisant le cadre théorique utilisé dans cette thèse pour modéliser les processus de contagion. On présente par la même occasion les formalismes compartimentaux, l'approche théorique privilégiée dans cette thèse. Ensuite, on introduit les *contagions simples* et les *contagions complexes*. Leur phénoménologie est illustrée à l'aide de quelques modèles typiques, grâce auxquels on définit plusieurs concepts importants et on établit une terminologie claire pour les chapitres à venir. Subséquemment, on montre comment la structure en réseau d'une population est intégrée dans les formalismes compartimentaux. Finalement, on discute de deux phénomènes exotiques : les transitions de phase *étendues* et la *localisation*.

### 2.1 Cadre théorique et formalismes compartimentaux

Le cadre théorique traditionnel des processus de contagion est celui des processus stochastiques avec un espace des états discret. Avec un ensemble d'états possible  $\mathcal{X}$ , on associe un état — une variable aléatoire —  $x_i \in \mathcal{X}$  à chaque individu  $i \in \mathcal{V}$  d'une population de taille  $|\mathcal{V}| = N$ . Les états les plus communément utilisés pour un individu sont : (S) *susceptible*, (I) *infecté* et (R) *retiré*. L'état susceptible indique que l'individu peut devenir infecté s'il est exposé à des individus infectés ; l'état infecté indique que l'individu est contagieux et peut transmettre la maladie ; l'état retiré implique que l'individu fut

infecté, a récupéré, et ne peut plus l'être en raison d'une immunité développée ou à la suite d'un décès. L'état du système complet est alors donné par le vecteur  $\mathbf{x} = [x_1, x_2, \dots, x_N] \in \mathcal{X}^N$ .

Pour modéliser l'évolution de l'état du système, autant les processus stochastiques en temps continu  $(\mathbf{x}(t))_{t \in \mathbb{R}^+}$  qu'en temps discret  $(\mathbf{x}(t))_{t \in \mathbb{N}_0}$  sont utilisés. On favorise la description en temps continu dans ce chapitre et pour la majorité de cette thèse (à l'exception du chapitre 5). Les unités du temps  $t$  et des taux de transition sont arbitraires, en fait, on travaille toujours avec des quantités sans unité grâce à l'adimensionnement des équations.

Dans cette thèse, on se concentre sur les processus de Markov homogène [98]; pour les processus non markoviens, le lecteur pourra se référer au chapitre 9 du livre de István Z. Kiss, Joel C. Miller et Péter L. Simon [163]. En temps continu, un processus de Markov homogène est formellement un processus de sauts (*jump process*) [98] dont l'évolution est régie par les *équations maîtresses*<sup>1</sup> suivantes :

$$\frac{d}{dt}P(\mathbf{x}, t) = \sum_{\mathbf{x}' \in \mathcal{X}^N} [w(\mathbf{x}' \rightarrow \mathbf{x})P(\mathbf{x}', t) - w(\mathbf{x} \rightarrow \mathbf{x}')P(\mathbf{x}, t)] , \quad (2.1)$$

où  $P(\mathbf{x}, t) \equiv P[\mathbf{x}(t) = \mathbf{x}]$  et  $w(\mathbf{x} \rightarrow \mathbf{x}')$  est le taux de transition de l'état  $\mathbf{x}$  vers l'état  $\mathbf{x}'$ .

Même si l'équation (2.1) est bien définie, le cardinal de l'espace des états est  $|\mathcal{X}|^N$ , donc résoudre l'équation maîtresse devient rapidement impossible d'un point de vue computationnel pour de grandes populations, à moins que le processus possède suffisamment de symétries permettant de réduire de manière considérable la dimension du système [163]. Par conséquent, il y a typiquement deux avenues à suivre : la simulation numérique du processus de contagion et les formalismes théoriques, qui permettent soit d'approximer le processus ou de résoudre exactement une version simplifiée. Dans cette thèse, on met l'accent sur l'approche théorique; le lecteur intéressé aux simulations numériques est invité à consulter l'annexe A.

On se concentre plus particulièrement sur les *formalismes compartimentaux*. L'idée générale de cette approche est de partitionner les individus de la population  $\mathcal{V}$  en *classes d'équivalence*  $v \subset \mathcal{V}$ , où chaque classe est définie par un ensemble d'attributs [19] : état de l'individu, facteurs socio-économiques, âge [206], etc. Lorsqu'on utilise une description réseau, on utilise également les propriétés associées à la structure pour définir les classes. Plus il y a d'attributs caractérisant les classes, plus la description mathématique est précise et fidèle au système, mais plus il devient difficile d'obtenir des résultats analytiques et de développer une intuition : un bon formalisme compartimental est donc l'art de choisir un bon compromis.

En définissant  $f(v, t)$  comme la fraction des noeuds appartenant à la classe  $v$  au temps  $t$ , un formalisme compartimental est défini à partir du système d'équations différentielles suivant :

$$\frac{d}{dt}f(v, t) = \sum_{v'} [w(v' \rightarrow v)f(v', t) - w(v \rightarrow v')f(v, t)] , \quad (2.2)$$

où  $w(v \rightarrow v')$  est le taux auquel un individu change de la classe  $v$  vers la classe  $v'$ . On note une forme semblable avec l'équation (2.1), mais avec une différence fondamentale : en général, le taux  $w(v' \rightarrow v)$  et la densité  $f(v, t)$  deviennent des variables aléatoires et le processus résultant est non markovien [71]. En revanche, dans la limite  $N \rightarrow \infty$ , on suppose que les fluctuations des densités et des taux à l'équation (2.2) sont négligeables et que le système d'équations converge vers un système déterministe.

1. En mathématique, on parle plus souvent des équations de Kolmogorov « vers l'avant ».

Cette correspondance est démontrée de manière rigoureuse notamment par les théorèmes de Thomas G. Kurtz [167, 168]. Par conséquent, dès lors que l'on considère un formalisme compartimental, la limite  $N \rightarrow \infty$  est toujours implicite. Notons toutefois qu'il est toujours possible d'effectuer un développement de van Kampen [98] pour approximer les fluctuations associées aux effets de taille finie, comme ce fut fait notamment dans le cadre des équations maîtresses approximatives [237].

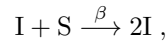
Pour compléter la définition d'un formalisme compartimental, il faut *estimer* les taux de transition  $w(v \rightarrow v')$  en fonction des classes d'équivalence et de la nature du processus de contagion. Dans la plupart des cas, le taux  $w(v \rightarrow v')$  ne sera pas exact, mais plutôt approximé grâce à certaines suppositions sur la structure des interactions.

Afin d'introduire les contagions simples et les contagions complexes, on utilise le formalisme compartimental le plus simple qui soit. On suppose que la population est de taille infinie, homogène (sans structure) et que les individus interagissent entre eux de manière uniforme. Cette approche correspond à l'approximation de *champ moyen* en physique statistique. Puisque les individus d'un même état sont alors *indistinguishables* du point de vue de la dynamique, on partitionne les individus uniquement en fonction de leur état. Pour les trois états introduits précédemment, on définit la fraction des individus susceptibles, infectés et retirés par  $S(t)$ ,  $I(t)$  et  $R(t)$  respectivement. On nomme également  $I(t)$  la *prévalence* de la maladie.

## 2.2 Contagions simples

On définit les contagions simples comme tout processus de contagion pour lequel la transmission de la maladie s'effectue de manière *indépendante* entre chaque paire d'individus.

Par exemple, on peut supposer que chaque individu entre en contact avec un autre individu (choisi aléatoirement) à un taux  $\beta$ . À la suite de ce contact, si l'un des deux individus est infecté et l'autre susceptible, ce dernier devient infecté. Ce processus est semblable à une réaction chimique de la forme



et c'est pourquoi la plupart des modèles de contagion simple suivent la *loi d'action de masse* [19] : le taux auquel de nouveaux individus deviennent infectés est directement proportionnel au produit de la fraction des individus infectés  $I(t)$  et de la fraction des individus susceptibles  $S(t)$ .

Une fois infecté, on suppose que chaque individu récupère à un taux  $\mu$ . Selon l'état de l'individu après avoir récupéré, on distingue deux modèles classiques.

### 2.2.1 Dynamique Susceptible-Infecté-Retiré (SIR)

Si l'individu devient retiré suite à sa récupération, cela correspond au modèle SIR. Celui-ci est régi par les équations différentielles suivantes :

$$\frac{dS}{dt} = -\beta IS , \tag{2.3a}$$

$$\frac{dI}{dt} = \beta IS - \mu I , \tag{2.3b}$$

$$\frac{dR}{dt} = \mu I . \tag{2.3c}$$

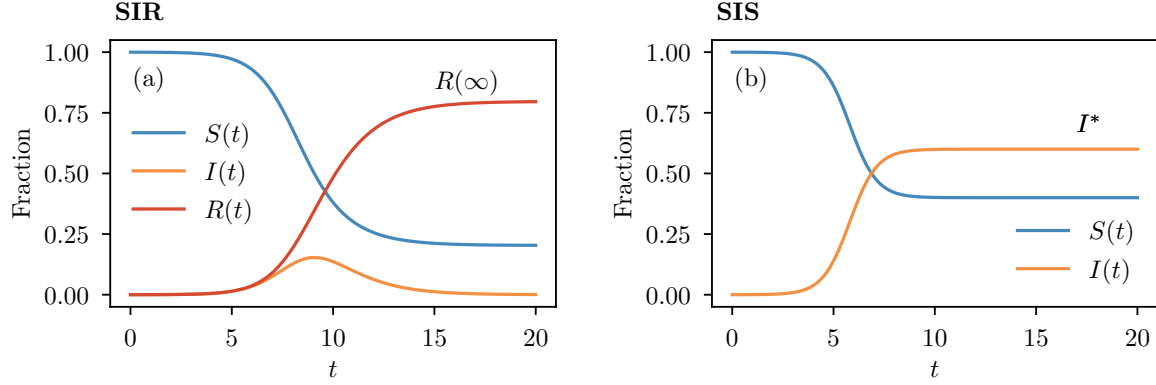


FIGURE 2.1 – Évolution temporelle pour les modèles SIR et SIS en champ moyen. Les systèmes sont initialisés avec une fraction d'individus infectés  $I(0) = 10^{-4}$ . (a)  $\beta = 2$ . Dans la limite  $t \rightarrow \infty$ , la fraction d'individus retirés converge vers la taille finale  $R(\infty)$ . (b)  $\beta = 2.5$ . Dans la limite  $t \rightarrow \infty$ , la fraction des individus infectés converge vers la prévalence stationnaire  $I^*$ .

L'une de ces équations est souvent omise en raison de la contrainte de normalisation,

$$S(t) + I(t) + R(t) = 1 . \quad (2.4)$$

On élimine un paramètre en rééchelonnant le temps en fonction du taux de récupération  $\mu$ , c'est-à-dire  $t \mapsto t\mu^{-1}$  et  $\beta \mapsto \beta\mu$ . On obtient alors

$$\frac{dS}{dt} = -\beta IS , \quad (2.5a)$$

$$\frac{dI}{dt} = \beta IS - I , \quad (2.5b)$$

$$\frac{dR}{dt} = I . \quad (2.5c)$$

On réutilise cette technique abondamment, ce qui est équivalent à fixer le taux de récupération à  $\mu = 1$ . Le taux  $\beta$  à l'équation 2.5 s'interprète désormais comme le *taux de reproduction de base*,  $\mathcal{R}_0$ , correspondant au nombre moyen de nouvelles infections causées par un nouvel individu infecté, lorsque la population est saine ( $S \rightarrow 1$ ) [12].

L'évolution temporelle typique du modèle SIR est présentée à la figure 2.1(a). Le nombre d'individus infectés croît, puis s'éteint, menant à une fraction finale d'individus ayant été atteints,  $R(t \rightarrow \infty) \equiv R(\infty)$ . La fraction  $R(\infty)$  est souvent nommée la *taille finale* de l'épidémie (bien qu'ici ce soit une fraction) ou le *taux d'attaque* [138]. Il est important de mentionner que  $R(\infty)$  dépend des conditions initiales, mais généralement on utilisera  $S(0) = 1 - \epsilon$ ,  $I(0) = \epsilon$  et  $R(0) = 0$  avec  $\epsilon \rightarrow 0$  pour représenter un seul *patient zéro* pour l'épidémie.

On illustre à la figure 2.2 le comportement de  $R(\infty)$  en fonction du paramètre  $\beta$ . Dans le langage de la physique statistique, on observe une transition de phase continue pour le paramètre d'ordre  $R(\infty)$  : si  $\beta \leq 1$ , alors  $R(\infty) \sim \epsilon \rightarrow 0$ , tandis que si  $\beta > 1$ , alors la taille finale de l'épidémie correspond à une fraction macroscopique de la population.

Pour caractériser analytiquement la transition de phase continue, on recherche la valeur critique  $\beta_c$  telle que pour  $\beta > \beta_c$ , la dérivée temporelle associée à  $I(t)$  est positive — le nombre d'individus infecté croît.

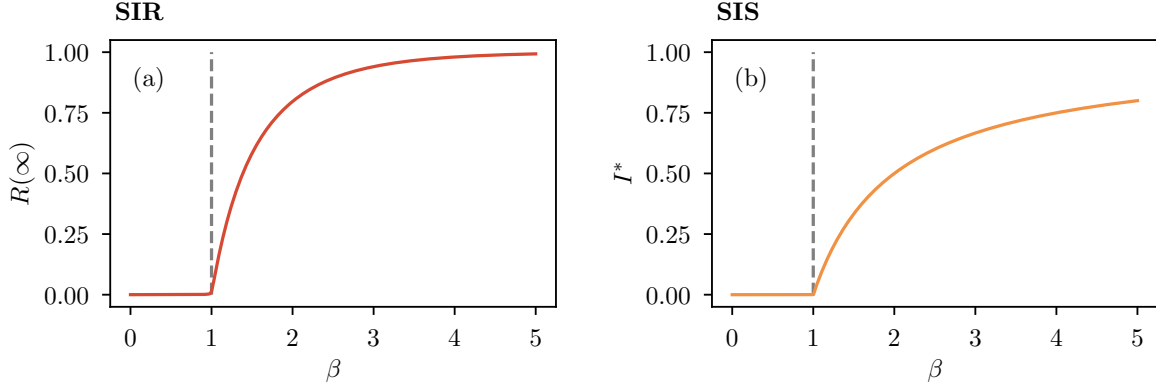


FIGURE 2.2 – Transition de phase des modèles SIR et SIS en champ moyen. Les systèmes sont initialisés avec une fraction d’individus infectés  $I(0) = 10^{-4}$ . (a) L’épidémie affecte une fraction macroscopique de la population si  $\beta > \beta_c = 1$ . (b) L’épidémie est endémique ( $I^* > 0$ ) si  $\beta > \beta_c = 1$ .

La quantité  $\beta_c$  est aussi communément appelée le *seuil d’épidémie*. En linéarisant l’équation (2.5b) pour  $\epsilon \rightarrow 0$ , on dégage

$$\frac{dI}{dt} \simeq (\beta - 1)I, \quad (2.6)$$

donc  $\beta_c = 1$ . De plus, une propriété très importante du système linéarisé est la suivante :

$$I(t) \sim e^{(\beta-1)t}. \quad (2.7)$$

Par conséquent, la croissance (ou décroissance) est exponentielle au début de l’épidémie. Ce résultat s’interprète de manière intuitive à l’aide du taux de reproduction de base : si  $\mathcal{R}_0 > 1$ , une personne infectée en infecte plus d’une en moyenne, donc le nombre d’individus infectés croît de manière exponentielle.

## 2.2.2 Dynamique Susceptible-Infecté-Susceptible (SIS)

Si l’individu devient susceptible suite à sa récupération, on obtient le modèle SIS. Celui-ci est régi par les équations différentielles suivantes :

$$\frac{dS}{dt} = -\beta IS + I, \quad (2.8a)$$

$$\frac{dI}{dt} = \beta IS - I, \quad (2.8b)$$

où on a substitué  $\mu = 1$ . En utilisant la contrainte de normalisation  $I(t) + S(t) = 1$ , on réduit le système à une seule équation,

$$\frac{dI}{dt} = \beta I(1 - I) - I. \quad (2.9)$$

Puisque les individus peuvent se faire infecter à nouveau après avoir récupéré, il est possible d’observer un état *endémique*, avec  $I^* > 0$  dans la limite  $t \rightarrow \infty$ , tel qu’illustrée à la figure 2.1(b).

À l’instar du modèle SIR, le modèle SIS possède une transition de phase continue. On utilise plutôt la fraction stationnaire  $I^*$  comme paramètre d’ordre. À la figure 2.2(b), on voit qu’il existe un état



$I^* > 0$  uniquement pour  $\beta > 1$ . On démontre ce résultat en posant le côté gauche de l'équation 2.9 comme étant égal à zéro. On trouve deux solutions,  $I^* = 0$  et

$$I^* = 1 - \frac{1}{\beta}. \quad (2.10)$$

Cette deuxième solution est positive uniquement pour  $\beta > 1$ . En linéarisant l'équation 2.9, on trouve que la solution triviale  $I^* = 0$  est stable uniquement pour  $\beta \leq 1$ ; pour  $\beta > 1$ , la solution positive  $I^* > 0$  devient stable. Dans le langage de la dynamique non linéaire, le système subit une bifurcation transcritique à  $\beta_c = 1$ , où les deux solutions s'échangent leur stabilité. Ces deux perspectives — transition de phase et bifurcation — sont utilisées de manière interchangeable dans cette thèse.

Au niveau de leur phénoménologie, les modèles SIR et SIS sont semblables et différents à la fois : les deux possèdent une transition de phase continue, mais la nature de celle-ci est très différente. Pour le premier, la dynamique est éteinte dans la limite  $t \rightarrow \infty$ , et on s'intéresse généralement à la fraction de la population qui est dans l'état retiré. Pour le deuxième, le système est dans un état d'équilibre dynamique avec une prévalence  $I^* > 0$  lorsque  $\beta > \beta_c$ . C'est pourquoi la solution  $I^* = 0$  est souvent nommée l'état *absorbant* ; même si cette solution est possiblement instable, le système reste figé dans cet état s'il est atteint. C'est surtout un aspect important à considérer pour des populations de taille finie et lors de la simulation numérique du processus.

Finalement, il existe plusieurs généralisations de ces deux modèles de contagion simple, le plus souvent formulées en ajoutant des états intermédiaires pour les individus, comme l'état *exposé* (E), pour lequel l'individu a contracté la maladie, mais n'est pas encore contagieux. Bien que les modèles plus sophistiqués du type Susceptible-Exposé-Infecté-Retiré (SEIR) ou Susceptible-Infecté-Retiré-Susceptible (SIRS) possèdent une dynamique plus riche, leur phénoménologie dans la limite  $t \rightarrow \infty$  se ramène aux deux cas déjà rencontrés.

Pour le reste de cette thèse, on met l'accent sur les processus pouvant mener à un état stationnaire  $I^* > 0$ , à l'instar du modèle SIS classique.

## 2.3 Contagions complexes

Notre définition des contagions complexes est simple une fois les contagions simples définies<sup>2</sup> — tout processus de contagion qui n'est pas une contagion simple est nécessairement une contagion complexe. Plus précisément, la contagion est complexe s'il est impossible de décomposer la transmission de la maladie en un ensemble de processus indépendants entre chaque paire d'individus.

Il est important de mentionner que cette interprétation des contagions complexes est très large et n'est pas adoptée par tous. La définition la plus couramment utilisée est plus restreinte : pour qu'un individu devienne infecté, il doit nécessairement être exposé à plusieurs individus infectés [52, 176]. En revanche, on rencontre aussi parfois une notion de seuil d'exposition qui est assoupli — on considère alors une probabilité d'infection qui varie de manière sigmoïdale avec le niveau d'exposition, plutôt que d'avoir un seuil fixe [207]. Ce type de processus possède une phénoménologie semblable aux contagions complexes plus orthodoxes, et c'est pourquoi on favorise une définition plus générale. Notre classification des contagions simples/complexes suit en fait une perspective équivalente à la classification des processus comme indépendants/interdépendants dans les articles de Dodds [74, 75].

---

2. Jeu de mots intentionnel.

### 2.3.1 Modèle de Granovetter

L’analogie entre la propagation des maladies infectieuses et certains phénomènes sociaux donna naissance au concept de contagion sociale. La propagation d’une rumeur [66, 67] ou la transmission d’une idée [108] furent étudiées sous cette perspective dès les années 1960. Cette analogie est même entrée dans le langage courant : une nouvelle ou une vidéo populaire partagée à grande échelle est souvent étiquetée comme « virale ».

Les premiers modèles de contagion sociale étaient des contagions simples, avec des fondements semblables aux modèles épidémiologiques. Par contre, ceux-ci ne reflètent pas adéquatement le processus de décision auquel un individu fait face, par exemple, dans le choix de joindre ou non une cause sociale, comme une révolte. Les individus pèsent le pour et le contre avant de faire leur choix, et la popularité du mouvement est certainement un aspect important à considérer : une révolte impopulaire a peu de chance de succès, et les conséquences associées pèsent plus lourdement sur le contre.

Le modèle de Granovetter s’inscrit dans cette perspective et considère que chaque individu d’une population a le choix de joindre ou non une cause sociale en fonction sa popularité. Chaque individu est donc soit infecté — a joint la cause — ou susceptible. Dans sa formulation originale, l’évolution s’effectue selon un processus itératif en temps discret. Au temps  $t + 1$ , chaque individu  $i$  remet en question son état actuel. Un individu susceptible devient infecté si la prévalence  $I(t)$  égale ou excède un certain seuil personnel  $K$ . À l’inverse, si la prévalence  $I(t)$  est en dessous du seuil personnel  $K$  d’un individu infecté, il devient susceptible. Puisque les gens n’ont pas tous la même tolérance au risque — certains sont plus audacieux, d’autres plus conservateurs — le modèle suppose que le seuil  $K$  est tiré d’une densité de probabilité  $\wp(K)$ . Par conséquent, l’évolution du système est régie par l’équation suivante :

$$I(t + 1) = \int_0^{I(t)} \wp(K) dK = P[K \leq I(t)] . \quad (2.11)$$

Puisque les individus basent leur décision sur l’état global de la population, c’est comme si tous les individus de la population étaient constamment en contact, donc on est toujours dans une perspective de champ moyen.

Afin d’avoir un cadre théorique cohérent avec le reste de ce chapitre, on transforme l’équation (2.11) en prenant la limite du temps continu :

$$\begin{aligned} I(t + 1) - I(t) &= -I(t) + P[K \leq I(t)] , \\ \frac{dI}{dt} &\approx -I + P(K \leq I) . \end{aligned} \quad (2.12)$$

Ce nouveau système s’interprète comme suit : à un taux unitaire, chaque individu remet en question son état actuel. La fraction  $|P(K \leq I) - I|$  correspond à ceux qui sont insatisfaits de leur état actuel, et le signe indique si la fraction  $I(t)$  va croître ou décroître. Notons que cette transformation n’affecte pas la position des points d’équilibre du système.

À la figure 2.3, on illustre une des propriétés importantes du modèle de Granovetter : l’émergence de plusieurs solutions stables dans la limite  $t \rightarrow \infty$ . En utilisant la distribution des seuils de la figure 2.3(a), il existe deux solutions stables,  $I^* = 0$  et  $I^* = 1$ , ayant chacune leur propre bassin d’attraction [figure 2.3(b)]. En d’autres mots, le système est dans une phase *bistable*. Ce phénomène contraste avec le modèle SIS classique, pour lequel une seule solution stable existe en tout temps.

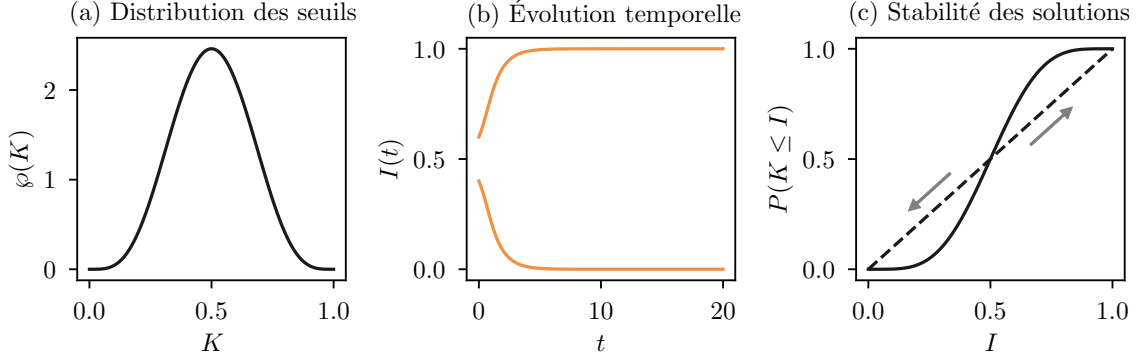


FIGURE 2.3 – Bistabilité émergente dans le modèle de Granovetter. (a) Distribution des seuils  $\varphi(K) \propto K^4(1-K)^4$ . (b) Évolution temporelle à partir de deux conditions initiales,  $I(0) = 0.4$  et  $I(0) = 0.6$ . (c) Les points fixes du système dynamique correspondent aux intersections entre la ligne continue  $P(K \leq I)$  et la ligne pointillée  $I$ . Lorsque la pente  $\partial_I P(K \leq I) < 1$ , le point fixe est stable, alors que si  $\partial_I P(K \leq I) > 1$ , le point fixe est instable.

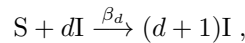
Les propriétés du système sont clarifiées à la figure 2.3(c). Les points fixes de l'équation (2.12) correspondent aux valeurs  $I^*$  satisfaisants  $I^* = P(K \leq I^*)$  — les intersections des courbes pointillées et pleines. Il y a trois points fixes, mais  $I^* = 0.5$  est instable : la direction du flot est illustrée par les flèches grises à la figure 2.3(c). En d'autres mots, pour envahir le système, une *masse critique* initiale est nécessaire, dans ce cas-ci  $I(0) > 0.5$ .

L'idée d'une masse critique est très importante pour les contagions sociales et correspond mieux à notre intuition pour certains phénomènes. Le modèle de Granovetter a conséquemment inspiré plusieurs travaux basés sur le concept de seuil [52, 75, 288]. Entre autres, Watts a transposé le modèle sur réseaux aléatoires [288]. Le seuil  $K$  ne s'applique plus sur la prévalence globale, mais plutôt localement sur la prévalence observée dans le voisinage d'un individu.

### 2.3.2 Contagion simpliciale

Un modèle plus récent de contagion complexe est celui de la contagion simpliciale [142]. L'article original introduit le modèle dans le contexte des réseaux d'ordre supérieur — plus particulièrement, des complexes simpliciaux —, mais la description analytique proposée utilise l'approximation de champ moyen.

Afin d'avoir une notation comparable à celle adoptée dans l'article original, on note  $d$  la dimension<sup>3</sup> d'une interaction impliquant  $n = d + 1$  individus. On suppose que chaque individu de la population participe à chaque interaction de dimension  $d$  à un taux  $\beta_d$ . Si un individu susceptible participe à une interaction de dimension  $d$  et que tous les autres individus y participant sont infectés, alors l'individu devient infecté. Comme pour le modèle SIS classique, chaque individu infecté récupère à un taux  $\mu = 1$ . L'augmentation du nombre d'individus infectés est encore une fois analogue à des réactions chimiques de la forme



3. Cela correspond à la dimension topologique des simplexes à  $n = d + 1$  sommets.

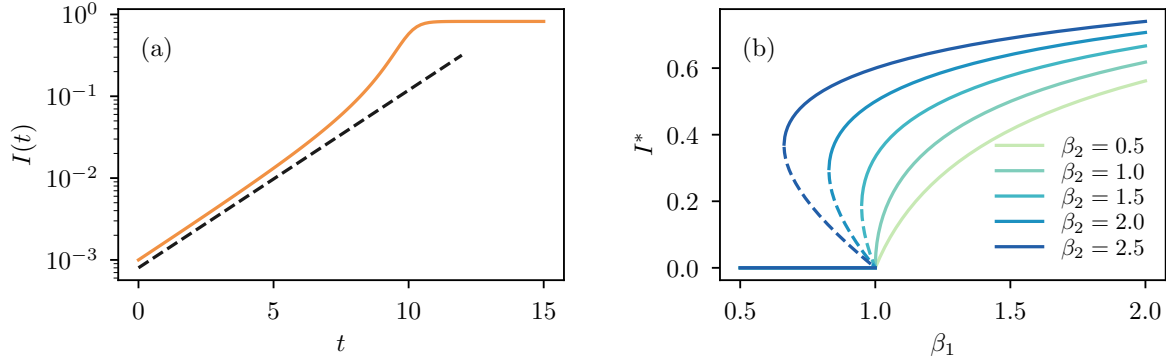


FIGURE 2.4 – Propriétés de la contagion simpliciale. (a) Croissance super-exponentielle de la prévalence  $I(t)$  avec  $\beta_1 = 1.5$  et  $\beta_2 = 5$ . (b) Transition de phase pour différentes valeurs  $\beta_2$ . Pour  $\beta_2 > 1$ , il y a émergence d’une phase bistable, avec deux solutions stables (lignes continues) et une solution instable (ligne pointillée).

auquel cas la loi d’action de masse s’applique. La prévalence évolue alors selon

$$\frac{dI}{dt} = -I + (1 - I) \sum_{d=1}^D \beta_d I^d, \quad (2.13)$$

où  $D$  est la dimension maximale des interactions.

Ce modèle combine à la fois la transmission entre paires d’individus ( $d = 1$ ) et la transmission par un groupe d’individus infectés ( $d > 1$ ), pour lequel le concept de seuil est implicite : il faut avoir  $d$  individus infectés dans l’interaction pour mener à une infection. Ce mécanisme de transmission permet entre autres de modéliser la pression sociale exercée par un groupe d’individus.

Par exemple, supposons un souper entre amis, lors duquel un certain nombre d’individus — infectés — partagent une opinion sur un sujet quelconque. Durant la soirée, un individu susceptible — qui n’a pas encore adopté cette opinion — pourra avoir des interactions dyadiques avec chacun des participants, mais il pourra aussi avoir des interactions (des discussions) de groupe, ce qui accroît la possibilité que l’individu se fasse convaincre. La fréquence et la force de ces différents types d’interaction sont encapsulées dans les taux  $\beta_d$ .

Afin d’analyser les propriétés du modèle, on se limite aux interactions de dimension  $D = 2$ , donc impliquant jusqu’à  $n = 3$  individus.

$$\frac{dI}{dt} = -I + (1 - I)(\beta_1 I + \beta_2 I^2). \quad (2.14)$$

La figure 2.4(a) présente l’évolution temporelle du processus pour une valeur  $\beta_2$  élevée. On remarque une différence importante par rapport aux contagions simples — la prévalence croît de manière *super-exponentielle*. Pour une prévalence faible, les interactions de dimension  $d = 2$  ont peu d’effet, car il est peu probable d’avoir une interaction avec deux individus infectés en même temps. Lorsque la prévalence devient suffisamment élevée, les interactions d’ordre supérieur contribuent de manière considérable, ce qui accélère la propagation de l’épidémie et mène à une croissance super-exponentielle.

Dans la limite stationnaire, il existe trois solutions possibles, à l’instar du modèle de Granovetter. Il y

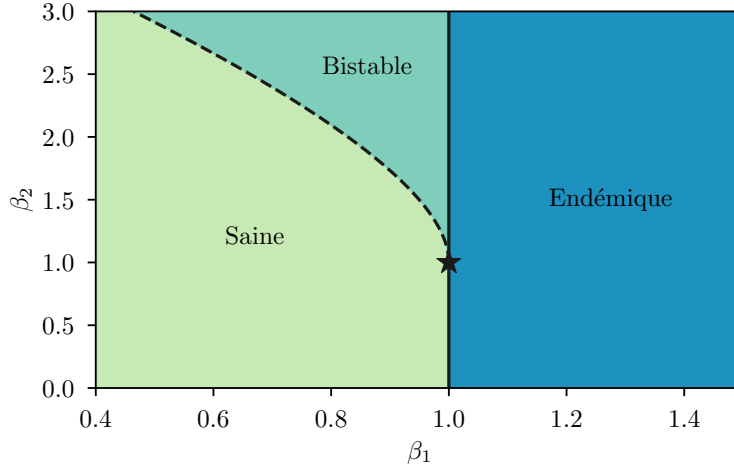


FIGURE 2.5 – Diagramme des phases du modèle de contagion simpliciale. La ligne critique pleine  $\beta_1 = \beta_1^c = 1$  est associée au seuil d’invasion. La ligne critique pointillée  $\beta_1 = \beta_1^p = 2\sqrt{\beta_2} - \beta_2$  pour  $\beta_2 > 2$  est associée au seuil de persistance. Le point tricritique  $(\beta_1, \beta_2) = (1, 1)$  identifié d’une étoile correspond au point où les deux lignes critiques se rejoignent.

a  $I^* = 0$  et

$$I_{\pm}^* = \frac{\beta_2 - \beta_1 \pm \sqrt{(\beta_1 - \beta_2)^2 - 4\beta_2(1 - \beta_1)}}{2\beta_2}. \quad (2.15)$$

En revanche, les deux solutions  $I_{\pm}^*$  ne sont pas réelles et positives pour tous les couples de paramètres  $(\beta_1, \beta_2)$ . En linéarisant l’équation (2.14), on trouve que  $I_+^*$  est stable alors que  $I_-^*$  est instable lorsque les deux solutions appartiennent à l’intervalle  $(0, 1]$ .

Comme illustré à la figure 2.4(b), si  $\beta_2 \leq 1$ , seulement la solution  $I_+^*$  est positive, et seulement pour  $\beta_1 > 1$ . Dans ce régime, la phénoménologie est identique à celle du modèle SIS classique — on a une transition de phase continue à  $\beta_1 = \beta_1^c = 1$ .

Si  $\beta_2 > 1$ , la solution  $I_-^*$  est positive sur l’intervalle  $[\beta_1^p, \beta_1^c)$ , tandis que la solution  $I_+^*$  est positive sur l’intervalle  $[\beta_1^p, \infty)$ . On nomme la quantité  $\beta_1^p = 2\sqrt{\beta_2} - \beta_2$  le *seuil de persistance*, car c’est la plus petite valeur permettant d’avoir une solution stationnaire  $I^* > 0$ . On nomme la quantité  $\beta_1^c$  le *seuil d’invasion*, car pour n’importe quelle condition initiale  $I(0) > 0$ , le système tend vers la solution stationnaire  $I_+^*$ . En fait, pour  $\beta_1 > \beta_1^c$ , la solution triviale  $I^* = 0$  devient instable grâce à une bifurcation transcritique, soit avec  $I_+^*$  si  $\beta_2 \leq 1$  ou avec  $I_-^*$  si  $\beta_2 > 1$ .

Dans l’intervalle  $[\beta_1^p, \beta_1^c]$ , le système est dans une phase bistable, à l’instar du modèle de Granovetter, mais cet intervalle existe seulement pour  $\beta_2 > 1$ . Pour cette raison, on nomme également la quantité  $\beta_2^c = 1$  le *seuil de bistabilité*. Il est alors plus commode d’analyser le diagramme des phases en deux dimensions dans l’espace  $(\beta_1, \beta_2)$ , comme illustré à la figure 2.5. Les lignes critiques pleine et pointillée sont associées aux seuils d’invasion et de persistance respectivement. Les deux lignes critiques séparent l’espace en trois phases distinctes : une phase saine, dans laquelle uniquement  $I^* = 0$  est stable, une phase bistable, et une phase endémique dans laquelle seulement la solution  $I_+^* > 0$  est stable. Le point  $(\beta_1^c, \beta_2^c) = (1, 1)$  correspond à un point *tricritique*, où les deux lignes critiques se rejoignent.

### 2.3.3 Force d'infection non linéaire

La *force d'infection* [129] correspond au nombre de nouveaux individus infectés par unité de temps, divisé par le nombre d'individus à risque (la population susceptible). En d'autres mots, cela correspond au taux d'infection pour chaque individu susceptible. Pour les modèles de contagion simple SIR et SIS, la force d'infection est simplement  $\beta I(t)$ , et varie donc *linéairement* avec la fraction d'individus infectés.

Pour les modèles de contagion complexe présentés dans cette section, on a plutôt une force d'infection *non linéaire*. En effet, pour le modèle de contagion simpliciale, la force d'infection est de la forme  $\sum_{d=1}^D \beta_d I^d$ . Pour le modèle de Granovetter, lorsque  $P(K \leq I) > I$ , le nombre d'individus infectés croît, auquel cas la force d'infection correspond à  $[P(K \leq I) - I]/(1 - I)$ . La non-linéarité mène à une la phénoménologie plus diverse pour les contagions complexes, notamment avec l'émergence d'une phase bistable [figure 2.3(b) et 2.4(b)] et la propagation super-exponentielle [figure 2.4(a)]. Plus généralement dans cette thèse, on associe les contagions complexes à une force (ou un taux) d'infection non linéaire.

Mentionnons que divers travaux ont déjà été réalisés sur les forces d'infection non linéaires [129, 182, 183]. Fait intéressant : les modèles SIRS et SEIRS peuvent avoir des solutions périodiques dans ce contexte. Un modèle qui fut étudié en détail est celui dans lequel la force d'infection a une forme en loi de puissance [182],  $\beta I^\nu$ , menant à l'équation suivante pour la version SIS :

$$\frac{dI}{dt} = -I + \beta(1 - I)I^\nu . \quad (2.16)$$

La solution triviale  $I^* = 0$  existe toujours, et il peut y avoir jusqu'à deux autres solutions non triviales. Lorsque  $\nu < 1$ , surprenamment, la solution triviale  $I^* = 0$  est toujours instable, et il y a une solution  $I^* > 0$  stable — en d'autres mots, il n'y pas de seuil d'invasion  $\beta_c$ . Pour  $\nu > 1$ , ce modèle possède une transition de phase discontinue, à l'instar du modèle de contagion simpliciale. Le seuil de persistance correspond à

$$\beta_p = \frac{\nu^\nu}{(\nu - 1)^{\nu-1}} , \quad (2.17)$$

mais il n'existe toujours pas de seuil d'invasion  $\beta_c$ . Il n'y a pas de phase endémique, seulement une phase saine ou bistable, ce qui requiert toujours une certaine masse critique initiale  $I(0)$  pour atteindre la solution stationnaire  $I^* > 0$ .

Comme ce sera montré aux chapitres 5 et 6, ces résultats ne tiennent plus lorsque le taux d'infection en loi de puissance est défini par rapport à la prévalence à l'intérieur des groupes. La granularité de la population et de la représentation réseau fait en sorte que l'on retrouve un seuil d'invasion  $\beta_c$  pour tout  $\nu$ . Le diagramme de phase est alors semblable à celui de la contagion simpliciale (figure 2.5).

## 2.4 Formalismes compartimentaux pour les contagions sur réseaux

Jusqu'à présent, les processus de contagion considérés se propageaient dans une population qui n'avait pas de structure : les individus pouvaient entrer en contact avec n'importe quel autre individu ou leur évolution était dictée selon l'état global du système  $I(t)$ . Les théories de champ moyen ont l'avantage d'être simples et mènent à une description mathématique de faible dimension, ce qui est commode pour développer une intuition analytique. Par contre, la structure des interactions sociales est beaucoup plus

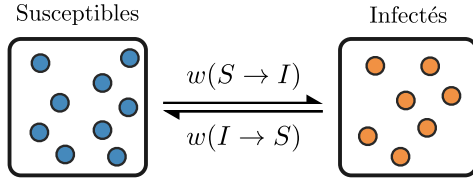


FIGURE 2.6 – Schéma du formalisme de champ moyen.

complexe qu’un « gaz » d’individus en interaction (voir chapitre 1). À la figure 2.6, on présente un schéma du formalisme compartimental utilisé dans les sections précédentes.

On suppose maintenant qu’il existe une structure réseau sous-jacente qui limite les interactions entre les individus de la population : les interactions sont limitées entre les individus qui sont *voisins* dans la représentation réseau du système. Il existe plusieurs approches théoriques permettant d’inclure la structure complète d’un système de taille finie, comme les théories de champ moyen centrées sur les individus, communément appelées *quenched mean-field theories* (QMF) [88, 231, 279] ou *microscopic markov chain approaches* [110, 196] (MMCA), ainsi que les généralisations s’intéressant aux liens entre les individus [194, 253] et aux groupes de noeuds [40].

Il y a aussi les approches par passage de messages (*message-passing*) [107, 149] et celles basées sur la percolation [5, 104, 159, 217, 288], qui s’appliquent aux processus pour lesquels les états évoluent de manière séquentielle ( $S \rightarrow I \rightarrow R$  par exemple). Il y eut aussi quelques tentatives pour appliquer ces techniques dans un contexte plus général [21, 228, 254].

Pour les processus ayant une dynamique des états qui est récurrente ( $S \rightarrow I \rightarrow S$  par exemple), les formalismes compartimentaux sont généralement les plus flexibles et les plus simples à analyser. Afin de généraliser la théorie aux populations définies par une structure réseau, il faut en revanche supposer que la structure sous-jacente est un réseau aléatoire. On s’intéresse particulièrement au modèle des configurations standard (pour les graphes) et au modèle des configurations biparti, tels que définis à la section 1.3.

#### 2.4.1 Théorie de champ moyen hétérogène

Le nombre de contacts potentiels entre un individu et l’ensemble de la population est sans contredit une propriété déterminante pour estimer le risque de contracter une maladie infectieuse. Pour un réseau d’interactions dyadiques (un graphe), le nombre de contacts est directement lié au degré  $k$  des noeuds, le tout encapsulé dans la distribution des degrés  $p_k$ . La manière la plus simple d’intégrer l’hétérogénéité sur le degré est d’utiliser une théorie de champ moyen hétérogène (*heterogeneous mean-field theory*; HMF) [37, 231–233]. Comme illustré à la figure 2.7, on partitionne les noeuds non plus seulement en fonction de leur état, mais également en fonction de leur degré. Un noeud susceptible (infecté) de degré  $k$  contribue à la fraction  $S_k(t)$  [ $I_k(t)$ ], satisfaisant la contrainte

$$\sum_{k=k_{\min}}^{k_{\max}} [S_k(t) + I_k(t)] = 1. \quad (2.18)$$

De plus, par définition,  $S_k(t) + I_k(t) = p_k$ , donc on se concentre sur la fraction  $I_k(t)$ .

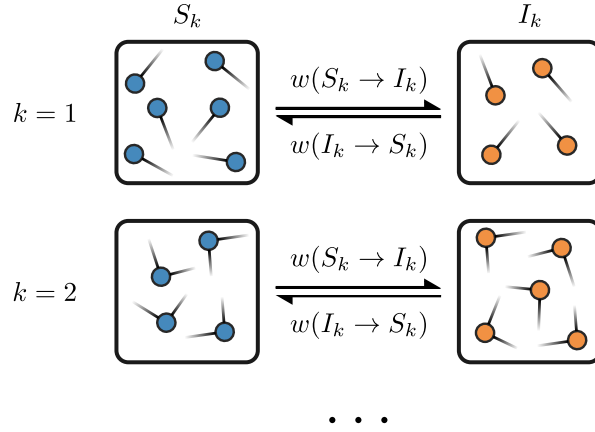


FIGURE 2.7 – Schéma du formalisme de champ moyen hétérogène.

L'évolution temporelle du système est décrite par

$$\frac{dI_k}{dt} = -\mu I_k + \Theta_k(1 - I_k), \quad (2.19)$$

où on suppose que les noeuds infectés récupèrent à un taux constant  $\mu$ , et que les noeuds susceptibles se font infectés en moyenne à un taux  $\Theta_k(t)$ . Jusqu'ici, aucune approximation n'a été faite (sauf l'hypothèse implicite que les équations sont valides dans la limite  $N \rightarrow \infty$ ). Toutefois, il faut maintenant définir le taux  $\Theta_k(t)$  en fonction de l'information disponible, soit  $\{I_k(t)\}_{k=k_{\min}}^{k_{\max}}$ .

Supposons le modèle SIS classique : chaque noeud infecté transmet la maladie à ses voisins à un taux  $\beta$ . Alors,

$$\Theta_k(t) = \beta k \theta_k(t), \quad (2.20)$$

où  $\theta_k(t)$  est la probabilité qu'un lien sortant d'un noeud susceptible de degré  $k$  mène à un noeud infecté. Il n'est pas possible d'exprimer  $\theta_k(t)$  de manière exacte en fonction des  $I_k(t)$ . Par conséquent, on utilise l'approximation suivante :

$$\theta_k(t) \approx \frac{1}{\langle k \rangle} \sum_k k I_k(t) \equiv \theta(t). \quad (2.21)$$

Cette approximation est constituée de deux volets :

1. on suppose que le graphe est tiré du modèle des configurations standard ;
2. on néglige complètement les *corrélations dynamiques*.

Par corrélations dynamiques, on sous-entend que la probabilité conjointe associée à l'état au temps  $t$  de deux noeuds  $i$  et  $j$ ,  $P(x_i, x_j)$ , est généralement différente du produit des marginales,  $P(x_i)P(x_j)$ . En ignorant ces corrélations, on utilise implicitement l'approximation  $P(x_i, x_j) \approx P(x_i)P(x_j)$ . Une autre manière d'interpréter l'approximation à l'équation (2.21) est que la structure du graphe est constamment en train de se réarranger de manière aléatoire (selon le modèle des configurations), et ce, à un taux beaucoup plus rapide que l'évolution de la dynamique sur le réseau — c'est le concept d'une structure *annealed*. En combinant les équations (2.19), (2.20) et (2.21), le système d'équations est fermé.



Les théories de champ moyen hétérogène peuvent être utilisées pour les contagions complexes également : l'approximation de la structure *annealed* fait en sorte que le nombre de noeuds infectés  $\ell$  dans le voisinage d'un noeud de degré  $k$  est tiré d'une distribution binomiale

$$B_{k,\ell}(t) = \binom{k}{\ell} \theta(t)^\ell [1 - \theta(t)]^{k-\ell}. \quad (2.22)$$

Si un noeud susceptible de degré  $k$  dont  $\ell$  voisins sont infectés devient infecté à un taux  $\beta(k, \ell)$ , le taux d'infection moyen pour un noeud susceptible de degré  $k$  est alors

$$\Theta_k(t) = \sum_{\ell=0}^k B_{k,\ell}(t) \beta(k, \ell). \quad (2.23)$$

Toutefois, pour une structure statique, ce type d'approximation n'est pas approprié pour la plupart des processus [106]. Les corrélations dynamiques deviennent trop importantes pour être négligées.

On remarque qu'en substituant un taux d'infection linéaire  $\beta(k, \ell) \mapsto \beta \ell$  dans l'équation (2.23), on retrouve le résultat de l'équation (2.20). En effet, comme c'était le cas pour la force d'infection dans les modèles sans structure réseau, les contagions simples sont associées à un taux d'infection *linéaire* et les contagions complexes sont associées à des taux d'infection *non linéaires*.

Il est aussi possible d'étendre la théorie en champ moyen hétérogène aux réseaux d'ordre supérieur. L'approche la plus naturelle est de partitionner les noeuds en fonction de leur appartenance et de leur état, comme nous le faisons au chapitre 5. D'autres approches furent proposées également [170] en mettant l'accent sur les groupes de taille  $n \in \{2, 3\}$ .

## 2.4.2 Équations maîtresses approximatives centrées sur les noeuds

Pour remédier au problème de la corrélation dynamique, les premières approches développées furent les fameuses approximations de paires (*pairwise approximations*) [155–157, 163]. Au lieu de partitionner les noeuds, on partitionne les liens selon l'état des noeuds aux extrémités, ou l'état et le degré des noeuds [78]. Ces approches décrivent adéquatement la corrélation dynamique entre chaque paire de noeuds connectés, mais ne permettent pas de saisir dans son ensemble les corrélations dynamiques dans le voisinage d'un noeud.

La solution : partitionner les noeuds non seulement en fonction de leur état et de leur degré, mais aussi en fonction de l'état de leurs voisins. Cette approche fut initialement introduite par Vincent Marceau *et al.* [189], puis reprise dans différents contextes [178, 274] et généralisée par James P. Gleeson [105, 106] et Peter G. Fennel [85] pour des dynamiques stochastiques arbitraires.

On définit  $S_{k,\ell}(t)$  [ $I_{k,\ell}(t)$ ] la fraction des noeuds qui sont susceptibles (infectés), de degré  $k$  et dont  $\ell \leq k$  des voisins sont infectés. Supposons que les noeuds infectés récupèrent toujours à un taux constant  $\mu$  et qu'un noeud susceptible devient infecter au taux  $\beta(k, \ell)$ . En prenant en compte tous les types de transitions possibles, comme à la figure 2.8, on écrit le système d'équations suivant :

$$\begin{aligned} \frac{dS_{k,\ell}}{dt} &= -\beta(k, \ell) S_{k,\ell} + \mu I_{k,\ell} - \Omega^S (k - \ell) S_{k,\ell} + \Omega_S (k - \ell + 1) S_{k,\ell-1} \\ &\quad - \mu \ell S_{k,\ell} + \mu (\ell + 1) S_{k,\ell+1}, \end{aligned} \quad (2.24a)$$

$$\begin{aligned} \frac{dI_{k,\ell}}{dt} &= \beta(k, \ell) S_{k,\ell} - \mu I_{k,\ell} - \Omega^I (k - \ell) I_{k,\ell} + \Omega_I (k - \ell + 1) I_{k,\ell-1} \\ &\quad - \mu \ell I_{k,\ell} + \mu (\ell + 1) I_{k,\ell+1}. \end{aligned} \quad (2.24b)$$

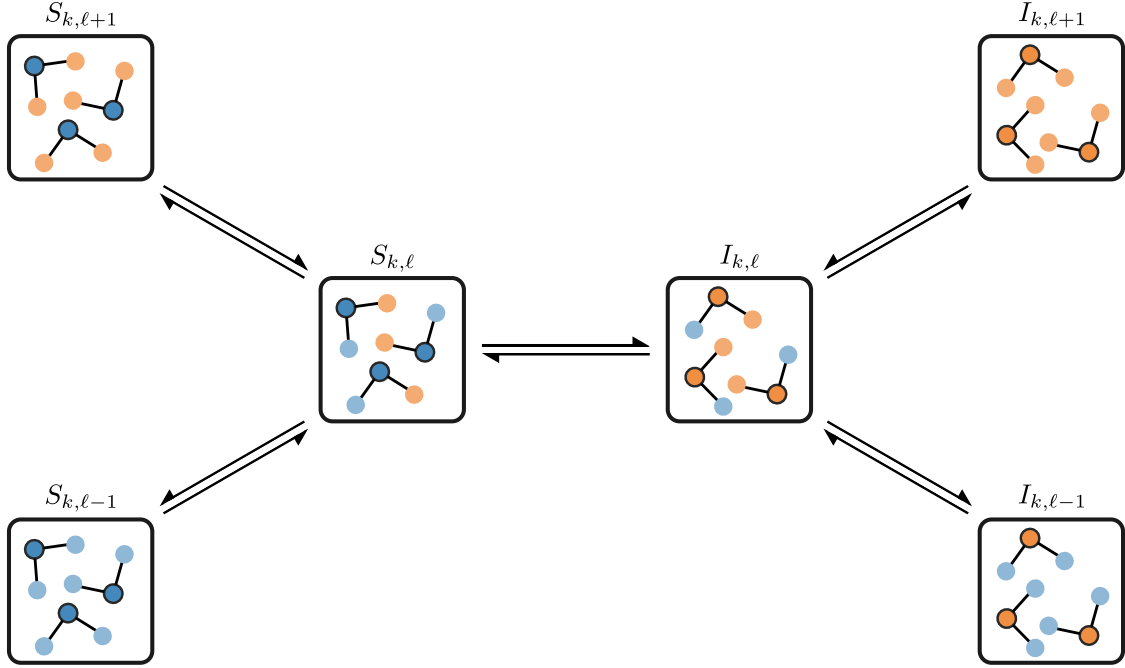


FIGURE 2.8 – Schéma du formalisme des équations maîtresses approximatives centrées sur les noeuds. Les taux entre les compartiments ne sont pas explicités afin d’alléger la figure.

Ce formalisme correspond à ce que l’on appelle des équations maîtresses approximatives (*approximate master equations* ; AMEs), dans ce cas-ci centrées sur les noeuds. On parle d’équations maîtresses, car les équations (2.24a) et (2.24b) apportent une description exacte des taux de transition pour l’état d’un noeud de degré  $k$  avec  $\ell$  voisins infectés. En revanche, les taux de transition associés aux voisins de ce noeud sont approximatés par  $\Omega_S(t)$  ou  $\Omega_I(t)$ , correspondant au taux d’infection moyen pour le voisin d’un noeud susceptible ou infecté respectivement.

Pour fermer le système, on doit estimer ces taux d’infection à partir des fractions  $S_{k,\ell}$  et  $I_{k,\ell}$ . En supposant encore une fois que le graphe sous-jacent est tiré du modèle des configurations standard, on utilise l’approximation suivante :

$$\Omega_S(t) = \frac{\sum_{k,\ell} \beta(k,\ell)(k-\ell)S_{k,\ell}}{\sum_{k,\ell} (k-\ell)S_{k,\ell}}, \quad (2.25a)$$

$$\Omega_I(t) = \frac{\sum_{k,\ell} \beta(k,\ell)\ell S_{k,\ell}}{\sum_{k,\ell} \ell S_{k,\ell}}. \quad (2.25b)$$

En effet, le voisin d’un noeud susceptible est également susceptible, de degré  $k$  et possède  $\ell$  voisins infectés avec une probabilité  $\propto (k-\ell)S_{k,\ell}$ , ce qui explique l’équation (2.25a). Un raisonnement similaire s’applique pour l’équation (2.25b). De ce fait, bien qu’on caractérise de manière précise les corrélations dynamiques de *courte portée* (dans le voisinage d’un noeud), on ignore les corrélations dynamiques de *longue portée*, c’est-à-dire tout ce qui est au-delà du voisinage d’un noeud. Par contre, les simulations numériques suggèrent que les corrélations à courte portée sont les plus importantes pour l’évolution temporelle des dynamiques sur graphes aléatoires [106].

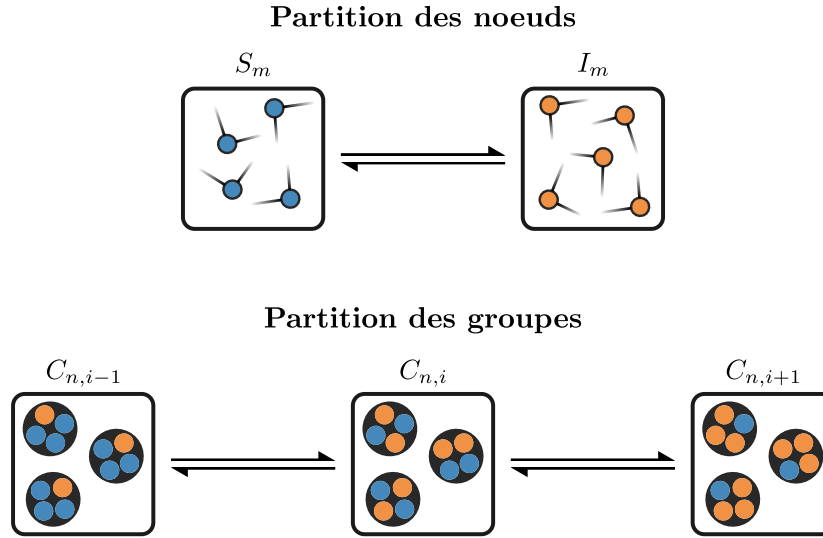


FIGURE 2.9 – Schéma du formalisme des équations maîtresses approximatives centrées sur les groupes. Les taux entre les compartiments ne sont pas explicités afin d’alléger la figure.

### 2.4.3 Équations maîtresses approximatives centrées sur les groupes

Pour résoudre les équations maîtresses centrées sur les noeuds [équation (2.24)], il faut intégrer  $O(k_{\max}^2)$  équations. De plus, le système d’équations laisse peu de place à une analyse théorique<sup>4</sup>. Si on tentait de généraliser l’approche aux réseaux d’ordre supérieur, force est d’admettre que le formalisme résultant serait encore plus complexe et opaque.

Dans le contexte des réseaux d’ordre supérieurs, si le processus de contagion est plutôt dominé par les *groupes* de noeuds, il n’est pas nécessaire d’avoir une description aussi détaillée qu’à l’équation (2.24) pour le voisinage des noeuds. Il serait, en revanche, avantageux d’avoir une description précise de l’état des groupes.

Pour ce faire, on partitionne les noeuds simplement en fonction de leur état et de leur appartenance (comme pour une théorie de champ moyen hétérogène), mais partitionner les groupes également, en fonction de leur taille et du nombre de noeuds infectés, comme à la figure 2.9. C’est en quelque sorte une généralisation des approximations de paires [155–157, 163] : au lieu de partitionner les liens, on partitionne des groupes de taille arbitraire. Ce type d’approche fut initialement introduit dans le contexte du modèle SIS classique — les groupes étaient des cliques ou des sous-graphes aléatoires [125]. Notons aussi qu’une approche semblable<sup>5</sup> avait été introduite pour la dynamique SIR afin modéliser l’influence des ménages sur la propagation [137].

On définit  $S_m(t)$  la fraction des noeuds qui sont susceptibles et d’appartenance  $m$  et  $C_{n,i}$  la fraction des groupes qui sont de taille  $n$  et dont  $i \leq n$  noeuds sont infectés. On se concentre sur la fraction susceptible des noeuds, car la fraction des noeuds infectés d’appartenance  $m$  correspond simplement à  $g_m - S_m$ . Supposons également une dynamique SIS standard comme dans la formulation originale [125],

4. Dans certains cas particuliers, il est possible de transformer les équations différentielles ordinaires en équations différentielles partielles sur des fonctions génératrices [255]. On peut alors utiliser la méthode des caractéristiques pour résoudre les équations.

5. L’approche ne considère toutefois pas une structure en réseau : la taille  $n$  des ménages est homogène et les individus interagissent selon un mécanisme d’action de masse à l’extérieur des ménages.

dans laquelle les groupes sont interprétés comme des cliques. Les noeuds infectés récupèrent à un taux  $\mu$  et transmettent la maladie à un taux  $\beta$ . L'évolution temporelle du système est alors régie par les équations suivantes :

$$\frac{dS_m}{dt} = \mu(g_m - S_m) - mrS_m, \quad (2.26a)$$

$$\begin{aligned} \frac{dC_{n,i}}{dt} = & \mu(i+1)C_{n,i+1} - \mu i C_{n,i} + (n-i+1)\{\beta(i-1) + \rho\}C_{n,i-1} \\ & - (n-i)\{\beta i + \rho\}C_{n,i}, \end{aligned} \quad (2.26b)$$

où  $r(t)$  est le taux d'infection moyen reçu de chaque appartenance à un groupe pour un noeud susceptible. Si on prend un noeud susceptible dans un groupe quelconque,  $\rho(t)$  correspond au taux d'infection moyen reçu de tous les *autres* groupes auxquels il appartient. Les équations (2.26a) et (2.26b) sont des équations maîtresses approximatives centrées sur les groupes, car la description des taux de transition à l'intérieur d'un groupe est exacte, mais le taux de transition associé aux appartenances à des groupes externes est approximé, dans ce cas-ci par le taux  $\rho(t)$ .

En supposant que la structure du réseau d'ordre supérieur provient du modèle des configurations biparti, on approxime les taux moyens par :

$$r(t) = \frac{\sum_{n,i} \beta i (n-i) C_{n,i}}{\sum_{n,i} (n-i) C_{n,i}}, \quad (2.27a)$$

$$\rho(t) = r(t) \frac{\sum_m m(m-1) S_m}{\sum_m m S_m}. \quad (2.27b)$$

En effet, un noeud susceptible est connecté à un groupe contenant  $n$  noeuds dont  $i$  sont infectés avec une probabilité  $\propto (n-i)C_{n,i}$ . Similairement, un noeud susceptible pigé aléatoirement dans un groupe possède une appartenance  $m$  avec une probabilité  $\propto mS_m$ .

Aux chapitres suivants, nous réutiliserons à plusieurs reprises ce formalisme, mais au lieu de travailler avec des fractions absolues  $S_m$  et  $C_{n,i}$ , nous utiliserons les probabilités conditionnelles  $s_m \equiv S_m/g_m$  et  $c_{n,i} \equiv C_{n,i}/p_n$ . De plus, comme ce sera montré au chapitre 6, il est possible d'étendre le formalisme aux contagions complexes et d'effectuer une analyse théorique du modèle en substituant le taux d'infection linéaire  $\beta i$  par une fonction générale — non linéaire — de la taille du groupe et du nombre de noeuds infectés,  $\beta i \mapsto \beta(n, i)$ .

## 2.5 Transition de phase étendue et localisation

Dans les sections précédentes, les transitions de phase continues que l'on a présentées étaient toutes *nettes*, au sens où près du point critique ( $\beta \rightarrow \beta_c^+$ ), le paramètre d'ordre, par exemple la prévalence globale  $I^*$ , se comportait comme

$$I^* \sim (\beta - \beta_c)^\xi, \quad (2.28)$$

avec un *exposant critique*  $\xi \leq 1$ , et  $I^* = 0$  pour  $\beta < \beta_c$ . Par exemple, pour le modèle SIS standard en champ moyen,  $\xi = 1$ . Pour la contagion simpliciale,  $I^* \sim (\beta_1 - \beta_1^c)$  pour  $\beta_1 \rightarrow \beta_1^+$  et  $\beta_2 < 1$ , mais  $I^* \sim (\beta_1 - \beta_1^c)^{1/2}$  pour  $\beta_2 = 1$ . Dans tous les cas, la transition de phase est continue, mais  $\partial_\beta I^*$  est discontinue au point critique, ce qui mène à une transition nette entre les deux phases, comme illustrée à la figure 2.10(a).

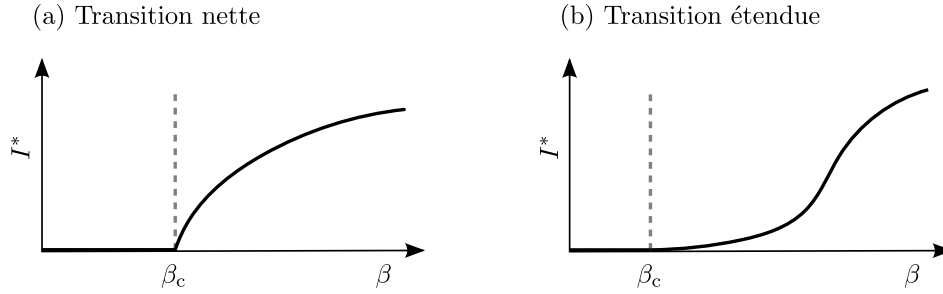


FIGURE 2.10 – Illustration d’une transition de phase nette et d’une transition de phase étendue.

En revanche, lorsqu’on considère des processus de contagion sur réseaux, la phénoménologie peut être altérée à un point tel que la transition de phase est *étendue* (*smearred*) [283]. Dans cette thèse, on dit que la transition de phase est étendue lorsque  $\partial_\beta I^*$  est continue et nulle au point critique. Par conséquent, l’exposant critique prend soit une valeur  $\xi > 1$ , ou le paramètre d’ordre peut ne pas avoir un comportement en loi de puissance, comme [283]

$$I^* \sim \exp[-a/(\beta - \beta_c)] , \quad (2.29)$$

pour  $\beta \rightarrow \beta_c^+$  et  $a > 0$ . La dérivée continue au point critique mène à une courbe en « S » du paramètre d’ordre, comme illustrée à la figure 2.10(b).

Les transitions de phase étendues interviennent souvent en physique statistique lorsque certains paramètres fixes sont des variables aléatoires — on parle alors de désordre « gelé » (*quenched disorder*). Pour un modèle d’Ising par exemple, à cause du désordre, les spins associés à différentes parties du système (des régions de taille finie) vont s’aligner de manière indépendante à des températures différentes en deçà de la température critique [283], ce qui détruit la transition de phase nette. Quant à elles, les transitions de phase étendues pour les processus de contagion sur réseaux sont souvent reliées à l’hétérogénéité de la structure [118] et à la *localisation* de l’épidémie près du point critique [62, 64, 223, 268].

La localisation est un phénomène qui intervient notamment en physique de la matière condensée, comme la localisation d’Anderson [11] : pour un niveau de désordre suffisant dans un réseau régulier, les électrons n’arrivent pas à se diffuser dans le milieu. Par analogie, on dira qu’une épidémie est localisée lorsque la contagion se propage principalement dans une portion restreinte du réseau, à laquelle on associe un sous-ensemble de noeuds  $\mathcal{L} \subset \mathcal{V}$ . Cela pourrait représenter une épidémie locale dans une communauté ou dans une portion vulnérable de la population — parmi les individus non vaccinés par exemple — par opposition à une contagion qui se propage dans l’entièreté de la population. Cette définition est intuitive, mais peu précise : dans une phase endémique, la probabilité qu’un noeud soit infecté est toujours non nulle si le réseau est connexe (une composante), ce qui laisse beaucoup de place à l’interprétation pour le régime de propagation localisé. Au chapitre 3, nous précisons les conditions pour lesquelles nous considérons une épidémie comme « localisée », et nous introduisons une mesure scalaire, le *ratio de participation inverse* [46, 109], pour caractériser la taille de l’ensemble  $\mathcal{L}$ .

Pour relier la localisation aux transitions de phase étendues, prenons l’exemple particulier de la localisation de la dynamique SIS standard sur graphes [45, 62, 64, 87, 91, 109, 223, 268]. Celle-ci est reliée à la localisation du plus grand vecteur de la matrice d’adjacence [46, 190, 229, 230, 252] à travers le formalisme QMF.

Pour des graphes issus du modèle des configurations avec une distribution des degrés en loi de puissance  $p_k \propto k^{-\gamma}$ , l'épidémie est localisée sur les noeuds de plus haut degré lorsque  $\gamma > 5/2$  [45, 109]. Pour  $\beta > \beta_c$ , l'épidémie est soutenue localement par les noeuds de degré  $k \geq k'$ , où  $k' \sim \beta^{-2}$  [268]. Par conséquent, chaque classe de degrés  $k$  possède un pseudo-seuil de transition  $\beta^{(k)} \sim 1/\sqrt{k}$  au-delà duquel les noeuds soutiennent la contagion dans leur voisinage de manière interdépendante. La prévalence  $I_k^*$  augmente de façon inhomogène pour les différentes classes de degré en une succession de pseudo-transitions [268], au lieu d'avoir une seule transition de phase nette pour l'ensemble des noeuds du système. Cela crée une transition de phase étendue, semblable à celle de la figure 2.10(b).

La localisation peut également affecter les processus de contagion possédant une transition de phase discontinue, bien qu'à notre connaissance il n'y ait aucune étude sur le sujet. Nous analysons et discutons de ce phénomène particulier à la section 6.6.

Deuxième partie

Contagions simples

## Chapitre 3

# Localisation mésoscopique des processus de contagion

Article original :

**Master equation analysis of mesoscopic localization in contagion dynamics on higher-order networks**

**Guillaume St-Onge**, Vincent Thibeault, Antoine Allard, Louis J. Dubé et Laurent Hébert-Dufresne

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

Référence : Phys. Rev. E **103**, 032301 (2021) [266]

© 2021 American Physical Society (§3.3-3.13)<sup>1</sup>

---

1. Ces sections contiennent le contenu original de l'article. Celui-ci n'a été modifié que pour se conformer au format exigé par la Faculté des études supérieures et postdoctorales de l'Université Laval.



### 3.1 Avant-propos

À la section 2.4.3, on a introduit les équations maîtresses approximatives centrées sur les groupes. Ce formalisme ouvre la voie à une description mathématique fiable des processus de contagion sur réseaux d'ordre supérieur et à l'analyse des conséquences liées à une structure hétérogène, autant au niveau de l'appartenance des noeuds qu'au niveau de la taille des groupes.

Dans ce chapitre, nous montrons que cette approche théorique peut être utilisée pour porter un regard nouveau sur le phénomène de localisation présenté à la section 2.5. Nous introduisons le concept de localisation *mésoscopique*. À l'instar de la localisation standard ou *microscopique*, l'épidémie est localisée sur un sous-ensemble des noeuds. Par contre, cette localisation sur les noeuds se trouve induite par une localisation au niveau mésoscopique — l'épidémie est localisée sur un sous-ensemble des *groupes*. Nous développons les outils théoriques pour analyser ce phénomène émergent et l'incidence sur la phénoménologie du système, notamment en exploitant les fonctions génératrices et l'analyse asymptotique.

Pour ce chapitre et le suivant (chapitres 3 et 4), nous nous limitons au cas des contagions simples. Celles-ci sont plus faciles à analyser mathématiquement, ce qui nous permet de développer une intuition analytique qui sera très utile plus tard lors de l'analyse des contagions complexes (chapitres 5 et 6). Puisqu'une contagion simple sur un hypergraphe est équivalente à une contagion simple sur un graphe dont les groupes sont des cliques, nous utilisons à quelques reprises cette correspondance pour obtenir nos résultats.

Symbole	Description
$n$	Taille de groupe : nombre de noeuds dans un groupe
$m$	Appartenance : nombre de groupes auxquels un noeud appartient
$n_{\max}$	Taille de groupe maximale
$m_{\max}$	Appartenance maximale
$\alpha$	Exposant reliant $m_{\max}$ et $n_{\max}$ par $m_{\max} \sim n_{\max}^{\alpha}$
$p_n$	Distribution des tailles de groupe
$g_m$	Distribution des appartenances
$\gamma_n$	Exposant de la loi de puissance $p_n \propto n^{-\gamma_n}$
$\gamma_m$	Exposant de la loi de puissance $g_m \propto m^{-\gamma_m}$
$i$	Nombre de noeuds infectés dans un groupe
$c_{n,i}(t)$	Fraction des groupes de taille $n$ possédant $i$ noeuds infectés au temps $t$
$s_m(t)$	Fraction des noeuds d'appartenance $m$ qui sont susceptibles au temps $t$
$I(t)$	Prévalence : fraction des noeuds qui sont infectés au temps $t$
$I_n(t)$	Fraction moyenne des noeuds qui sont infectés dans un groupe de taille $n$ au temps $t$

$\beta$	Taux auquel un noeud infecté transmet la maladie à ses voisins
$\beta_c$	Seuil d'épidémie : $I^* > 0$ existe pour $\beta > \beta_c$
$r(t)$	Taux d'infection moyen par groupe auquel un noeud susceptible est exposé au temps $t$
$\rho(t)$	Taux d'infection moyen pour tous les <i>autres</i> groupes auquel un noeud susceptible est exposé au temps $t$ , si on choisit le noeud aléatoirement dans un groupe
$h_{n,i}(\beta)$	Dérivée de $c_{n,i}^*$ par rapport à $\rho^*$ lorsque $\rho^* \rightarrow 0^+$
$H_n(x; \beta)$	Fonction génératrice de la suite $(h_{n,i})_{i=0}^n$
$\Omega(g_m, p_n)$	Couplage entre les groupes
$\tilde{Y}_4$	Ratio de participation inverse pour les noeuds
$\hat{Y}_4$	Ratio de participation inverse pour les groupes

TABLEAU 3.1 – Glossaire des symboles importants utilisés aux chapitres 3 et 4.

### 3.2 Résumé

Les modèles standards pour la propagation des maladies infectieuses supposent généralement des contacts aléatoires entre les individus d'une population. Toutefois, la réalité est bien plus complexe qu'un ensemble de rencontres entre paires d'individus : les interactions se produisent surtout lors de rassemblements, comme au travail, à l'intérieur des ménages, à l'école, pendant un concert, etc. La meilleure manière de prendre en compte ce type de rassemblements est par un réseau d'ordre supérieur. Dans cet article, nous modélisons les processus de contagions sur réseaux d'ordre supérieur à l'aide d'équations maîtresses approximatives. Nous décrivons exactement l'évolution de l'état pour les groupes de noeuds, que nous mettons en relation par une approximation de champ moyen. En étudiant la dynamique Susceptible-Infecté-Susceptible, notre approche révèle l'existence d'un régime de localisation mésoscopique, pour lequel le processus de contagion persiste presque exclusivement à l'intérieur des groupes de grande taille. La transition de phase de ce régime est d'ailleurs étendue sur un sous-domaine du paramètre de transmission, en raison de l'activation inhomogène des groupes. Au niveau mésoscopique, la distribution pour le nombre de noeuds infectés à l'intérieur des groupes d'une même taille peut être très dispersée, parfois même bimodale. Pour des structures hétérogènes, à la fois pour l'appartenance des noeuds et la taille des groupes, nous délimitons analytiquement la région de l'espace des paramètres structurels associé à la localisation mésoscopique. Nous mettons en relation ce phénomène avec celui de la localisation du vecteur propre dominant de la matrice d'adjacence. Cela nous permet de mettre en évidence l'importance de la représentation d'ordre supérieur afin de détecter la localisation mésoscopique. Finalement, nous discutons de l'influence de ce phénomène sur les interventions structurelles, ce qui permet d'apporter un regard nouveau sur un éventail de processus dynamiques.

### 3.3 Abstract

Simple models of infectious diseases tend to assume random mixing of individuals, but real interactions are not random pairwise encounters: they occur within various types of gatherings such as workplaces, households, schools, and concerts, best described by a higher-order network structure. We model contagions on higher-order networks using group-based approximate master equations, in which we track all states and interactions within a group of nodes and assume a mean-field coupling between them. Using the susceptible-infected-susceptible dynamics, our approach reveals the existence of a mesoscopic localization regime, where a disease can concentrate and self-sustain only around large groups in the network overall organization. In this regime, the phase transition is smeared, characterized by an inhomogeneous activation of the groups. At the mesoscopic level, we observe that the distribution of infected nodes within groups of the same size can be very dispersed, even bimodal. When considering heterogeneous networks, both at the level of nodes and at the level of groups, we characterize analytically the region associated with mesoscopic localization in the structural parameter space. We put in perspective this phenomenon with eigenvector localization and discuss how a focus on higher-order structures is needed to discern the more subtle localization at the mesoscopic level. Finally, we discuss how mesoscopic localization affects the response to structural interventions and how this framework could provide important insights for a broad range of dynamics.

### 3.4 Introduction

Classic epidemiological models have been successful at providing meaningful insights on the spreading of infectious diseases [12, 72]. Their simplicity is their strength : from as little information as the basic reproduction number  $\mathcal{R}_0$ , one can tell whether or not a disease should invade or not a population. However, we cannot hope to represent the complexity of human behavior and of our modern social structure with mathematical models relying solely on an average individual. This is even more true when considering more complex types of spreading processes, such as social contagions [52, 176, 207] or the coevolution of diseases [122].

The study of spreading processes on networks allows us to look beyond the mass-action principle, to account for more realistic contact patterns while keeping our models simple enough to provide meaningful insights [163, 231]. One success of network science has been to unveil the impact of the heterogeneity of contacts, and how this affects critical properties of these systems. Heterogeneous mean-field theories [37, 232], heterogeneous pair approximations [78, 193], and approximate master equations [105, 178, 189] represent only a few of the many techniques that have been developed to describe the behavior of dynamical processes on networks with heterogeneous number of contacts.

Social networks, however, are more than just random contacts between heterogeneous individuals : interactions occur in a coordinated manner because of a higher-level organization. At the mesoscopic level, we see groups of individuals that are more or less densely connected to one another [103, 218]. We can thus shift from asking *if* a contagion can invade a population, to *where* it should thrive within that population. This question is best embodied by the phenomenon of *epidemic localization* : near the epidemic threshold, the disease exists only in some parts of the whole network.

The localization of epidemics has been studied mostly through the lens of extensive numerical simulations or quenched mean-field theory [45, 109, 180, 230]. A general observation is that for most complex

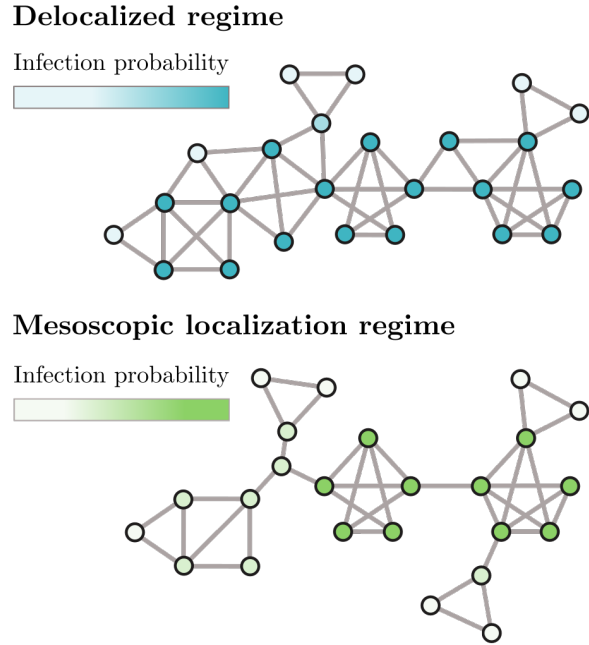


FIGURE 3.1 – Simple illustration of the mesoscopic localization phenomenon. In both regimes, the contagion is concentrated around the innermost core of the network, but the composition of the core is different. In what we called the delocalized regime, substructures of all sizes (e.g., triangles, 4–cliques, etc.) contribute to the contagion, while in the mesoscopic localization regime, there is a strong bias toward the largest and densest substructures.

networks, an epidemic should be localized around either the innermost network core or the hubs [45]. The localization subgraph depends on the structure, but also on the details of the dynamics [63, 91]. Moreover, localization dramatically affects the fundamental critical properties of an epidemic : it is notably possible to observe a Griffiths phase, where the system slowly relaxes to an inactive state [62, 64, 223, 224, 283]. Another notable effect is the smearing of the phase transition, where the order parameter develops inhomogeneously beyond the critical point [62, 118, 223, 268, 283].

Despite the important body of work on epidemic localization, theoretical results are still limited to a handful of models and most works have used a node-centric perspective. To broaden our understanding of localization of dynamical processes and embrace the higher-level organization of complex networks, we argue that higher-order representations of networks should be used [22]. Furthermore, we claim that approximate master equations represent powerful and flexible approaches for this purpose.

In this chapter, we reveal a phenomenon we call *mesoscopic localization* for spreading processes on certain heterogeneous networks with a higher-level organization (see Fig. 3.1). It is characterized by the localization of the contagion in large but finite-size mesoscopic substructures near the epidemic threshold<sup>2</sup>, with a phase transition that is smeared at the global level. To illustrate this phenomenon, we use a group-based framework together with an approximate master equation analysis of the susceptible-infected-susceptible (SIS) dynamics. We present a complete analytical description of the mesoscopic localization regime, while we describe its impact on interventions in chapter 4 to show how

<sup>2</sup>. Substructures at the mesoscopic level are meant to describe a large number of nodes, organized in more or less compact groups, but are by nature nonextensive. They could stand for schools, households, sports teams, etc. in a contact network.

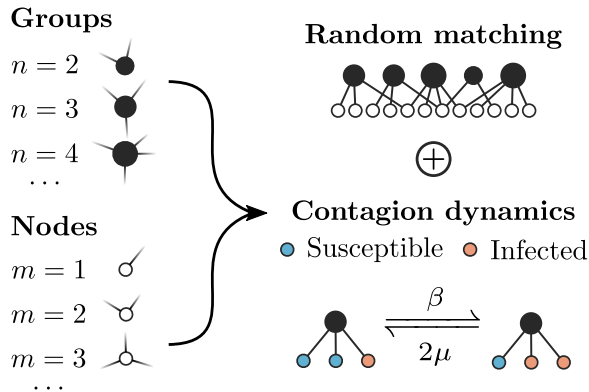


FIGURE 3.2 – Framework for contagions on higher-order networks. Nodes are assigned to  $m$  groups and groups are of various sizes  $n$ , distributed according to  $g_m$  and  $p_n$ . Groups are equivalent to cliques in the main text. We consider a SIS dynamics where infected nodes transmit the disease at rate  $\beta$ , and recover at rate  $\mu$ .

accounting for this localization regime is critical to our ability to control contagions on networks.

This chapter is structured as follows. First, we introduce the group-based framework and the approximate master equations in Sec. 3.5. We obtain an implicit expression and explicit bounds for the epidemic threshold in Sec. 3.5.1. With a development of the stationary state near the critical point, we show in Sec. 3.5.2 that mesoscopic localization emerges from a sufficiently weak coupling between the groups. Second, we fully characterize mesoscopic localization in Sec. 3.6. We derive asymptotic results for the scaling of the epidemic threshold in Sec. 3.6.1, leading to explicit expressions for the localization regimes. We then consider the effects of finite-size cutoffs in Sec. 3.6.2. We complete our analysis using the *inverse participation ratio*, further connecting our work with the literature on eigenvector localization. Our comparison reveals the importance of a change of perspective—a focus on higher-level group organization rather than individual nodes—in order to detect localization phases at the mesoscopic level. Finally, in Sec. 3.7, we discuss possible extensions of our work and some direct implications for the control of epidemics [267].

### 3.5 Group-based SIS model

There exists multiple representations for higher-order structures [22], ranging from simplicial complexes to hypergraphs, or more simply a bipartite graph, where nodes are attached to *groups* that encode the interaction. In this chapter, we use the latter [see Fig. 3.2]. Groups could be used to represent any kind of mesoscopic substructures, dense or sparse, with possibly weighted and directed edges. They could also be used to model higher-order interactions that cannot be decomposed into simpler pairwise interactions. To simplify the mathematical description, in the main text we consider that all groups of  $n$  nodes represent *cliques*, i.e., fully connected and undirected subgraphs. In Appendix 3.12, we generalize the approach to consider weighted cliques.

We assume that each node in the network belongs to a certain number of groups,  $m$ , the *membership* of the node, which is drawn from a membership distribution  $g_m$ . The size  $n$  of a group is drawn from a group size distribution  $p_n$ . We consider infinite-size heterogeneous random networks where nodes are assigned to groups uniformly at random [218]. In other words,  $m$  and  $n$  are uncorrelated. Throughout

the chapter, we denote expected values taken over  $p_n$  and  $g_m$  as  $\langle \dots \rangle$ , where the interior of the bracket makes it clear over which distribution the average is performed.

Let us introduce a few structural properties associated with this ensemble. The average membership of a node is  $\langle m \rangle$  and the average group size is  $\langle n \rangle$ . If we pick a node at random and follow a group to which it belongs, the distribution for the size of that group is proportional to  $np_n$ . Consequently, the average *excess* group size, i.e., the average number of neighbors this node has in that group, is  $\langle n(n-1) \rangle / \langle n \rangle$ . Since  $m$  and  $n$  are uncorrelated, the average degree of a node (in the one-mode projection of the bipartite graph) is therefore

$$\frac{\langle m \rangle \langle n(n-1) \rangle}{\langle n \rangle}.$$

On these networks, we consider the SIS dynamics in which each node is either infected or susceptible. Infected nodes transmit the disease to their neighbors at rate  $\beta$  and recover to the susceptible state at a rate  $\mu$  [see Fig. 3.2].

We describe the dynamics using the heterogeneous clique approximation of Ref. [125]. We track  $s_m(t)$ , the probability for a node of membership  $m$  to be susceptible at time  $t$ , and  $c_{n,i}(t)$ , the probability to observe  $i$  infected nodes within a group of size  $n$  at time  $t$ .

We define the following system of approximate master equations

$$\frac{ds_m}{dt} = \mu(1 - s_m) - mrs_m, \quad (3.1a)$$

$$\begin{aligned} \frac{dc_{n,i}}{dt} = & \mu(i+1)c_{n,i+1} - \mu ic_{n,i} + (n-i+1)\{\beta(i-1) + \rho\}c_{n,i-1} \\ & - (n-i)\{\beta i + \rho\}c_{n,i}, \end{aligned} \quad (3.1b)$$

which contains a total of  $O(m_{\max} + n_{\max}^2)$  equations, where  $m_{\max}$  and  $n_{\max}$  are the maximal membership and maximal group size respectively. From now on, we set  $\mu \equiv 1$  without loss of generality.

The mean fields  $r(t)$  and  $\rho(t)$  are defined as

$$r(t) = \frac{\sum_{n,i} \beta i (n-i) c_{n,i}(t) p_n}{\sum_{n,i} (n-i) c_{n,i}(t) p_n}, \quad (3.2a)$$

$$\rho(t) = r(t) \left[ \frac{\sum_m (m-1) m s_m(t) g_m}{\sum_m m s_m(t) g_m} \right]. \quad (3.2b)$$

If we take a susceptible node and select a random group to which it belongs,  $r(t)$  is the mean infection rate associated to that group. Indeed, the joint distribution for the size  $n$  of the group and the number of infected nodes  $i$  within that group is proportional to  $(n-i)c_{n,i}(t)p_n$ . Then  $r(t)$  is just an average of the infection rate received,  $\beta i$ , over this joint distribution.

Now if we pick a susceptible node in a group,  $\rho(t)$  is the mean infection rate received from all external groups (i.e., excluding the one we picked the node from). Assuming that infection coming from different groups are independent processes, we multiply  $r(t)$  with the mean *excess* membership of a susceptible node to get  $\rho(t)$ . The membership distribution of a susceptible node picked in a group is proportional to  $ms_m(t)g_m$ , thus we simply average  $m-1$ , the excess membership, over this distribution.

The global prevalence (average fraction of infected nodes) is

$$I(t) = \sum_m g_m [1 - s_m(t)] , \quad (3.3)$$

and the prevalence within groups of size  $n$  is

$$I_n(t) = \sum_i \frac{i}{n} c_{n,i}(t) . \quad (3.4)$$

Note that unless specified otherwise, sums over  $m$  ( $n$ ) are over every value such that  $g_m > 0$  ( $p_n > 0$ ), and sums over  $i$  cover the range  $\{0, \dots, n\}$ .

In Eq. (3.1), the evolution of each  $s_m$  is treated in a mean-field fashion<sup>3</sup>, while the evolution of each  $c_{n,i}$  is described using a master equation. The infection rate due to infected nodes *within* a group is treated exactly, while the contribution of infected nodes in *external* groups is approximated (i.e., the terms involving  $\rho$ ). We therefore refer to our approach as *approximate master equations*.

The system eventually settles to a stationary state in the limit  $t \rightarrow \infty$ , and henceforth we assume that the quantities  $s_m, c_{n,i}, r$  and  $\rho$  have reached a fixed point. These variables characterizing the stationary state are obtained by solving the following self-consistent expressions

$$s_m = \frac{1}{1 + mr} , \quad (3.5a)$$

$$(i+1)c_{n,i+1} = \{i + (n-i)[\beta i + \rho]\}c_{n,i} - (n-i+1)[\beta(i-1) + \rho]c_{n,i-1} , \quad (3.5b)$$

obtained from Eq. (3.1), and where  $r$  and  $\rho$  are still given by Eq. (3.2). It will be useful to rewrite Eq. (3.5b) more explicitly as

$$c_{n,i} = c_{n,0} \frac{n!}{(n-i)!i!} \prod_{j=0}^{i-1} [\beta j + \rho] \quad \forall i \in \{1, \dots, n\} , \quad (3.6)$$

with  $c_{n,0} = 1 - \sum_{i=1}^n c_{n,i}$ .

### 3.5.1 Epidemic threshold

For the SIS dynamics, there exists a critical value  $\beta_c$  for the transmission rate, called the epidemic threshold. For  $\beta < \beta_c$ , the absorbing state—where all nodes are susceptible—is attractive for all initial conditions. For  $\beta > \beta_c$ , the absorbing state becomes unstable and there exists a nontrivial stationary state.

To obtain an expression for  $\beta_c$ , let us redefine the stationary state observables as functions of  $\rho$ , i.e.,  $r(\rho)$ ,  $s_m(\rho)$  and  $c_{n,i}(\rho)$ . We then define the right-hand side of Eq. (3.2b) as  $F(\rho)$ . Since  $F(\rho)$  is bounded from above<sup>4</sup>, a positive solution  $\rho = F(\rho)$  exists if

$$\left. \frac{dF}{d\rho} \right|_{\rho \rightarrow 0} > 1 .$$

---

3. Note that we are still preserving the dynamic correlations between pairs of nodes by tracking each  $c_{n,i}$ .

4. In Eqs. (3.2a-b),  $c_{n,i}$  and  $s_m$  remain bounded for all  $\rho$ .

At the epidemic threshold  $\beta_c$ , this derivative is exactly 1, resulting in  $\rho \rightarrow 0$ ,  $r(\rho) \rightarrow 0$ ,  $s_m(\rho) \rightarrow 1$  and  $c_{n,i}(\rho) \rightarrow \delta_{i,0}$ , where  $\delta_{i,j}$  is the Kronecker delta.

It will prove useful to expand  $c_{n,i}$  near the epidemic threshold as  $c_{n,i}(\rho) = \delta_{i,0} + h_{n,i}\rho + O(\rho^2)$ . From Eq. (3.6), we obtain

$$h_{n,i}(\beta) \equiv \left. \frac{dc_{n,i}}{d\rho} \right|_{\rho \rightarrow 0} = \frac{n!\beta^{i-1}(i-1)!}{(n-i)!i!} \quad \forall i \in \{1, \dots, n\}, \quad (3.7)$$

and by definition  $h_{n,0} \equiv -\sum_{i=1}^n h_{n,i}$ .

For all  $n$ , we encode each sequence  $(h_{n,i})_{i=0}^n$  in the generating function

$$\begin{aligned} H_n(x; \beta) &= \sum_i h_{n,i} x^i, \\ &= h_{n,0} + \frac{1}{\beta} \sum_{i=1}^n \frac{n!}{(n-i)!i!} (\beta x)^i (i-1)!, \\ &= h_{n,0} + \frac{1}{\beta} \int_0^\infty \sum_{i=1}^n \frac{n!}{(n-i)!i!} (\beta u x)^i u^{-1} e^{-u} du, \\ &= h_{n,0} + \frac{1}{\beta} \int_0^\infty [(1 + \beta u x)^n - 1] u^{-1} e^{-u} du. \end{aligned} \quad (3.8)$$

Interestingly, the auxiliary generating function

$$Q_n(x; \beta) = \frac{H_n(x; \beta) - h_{n,0}}{H_n(1; \beta) - h_{n,0}} = \frac{\int_0^\infty [(1 + \beta u x)^n - 1] u^{-1} e^{-u} du}{\int_0^\infty [(1 + \beta u)^n - 1] u^{-1} e^{-u} du},$$

can be interpreted as the probability generating function for the *quasistationary distribution* (only for  $i > 0$ ) of the number of infected nodes in a group of size  $n$ , under the influence of a weak (vanishing) external field.

These generating functions allow us to write

$$\left. \frac{dF}{d\rho} \right|_{\rho \rightarrow 0} = \beta \frac{\langle m(m-1) \rangle}{\langle m \rangle \langle n \rangle} \langle (n-1)H'_n(1; \beta) - H''_n(1; \beta) \rangle, \quad (3.9)$$

where the derivatives are with respect to  $x$  and we have used standard properties of generating functions in combination with Eqs. (3.2a) and (3.2b). We simplify the above equation by noting that

$$(n-1)H'_n(1; \beta) - H''_n(1; \beta) = n(n-1) \int_0^\infty (1 + \beta u)^{n-2} e^{-u} du.$$

The epidemic threshold  $\beta_c$  is thus obtained by solving the following implicit equation for  $\beta$

$$\beta \frac{\langle m(m-1) \rangle}{\langle m \rangle \langle n \rangle} \langle n(n-1)A_n(\beta) \rangle = 1, \quad (3.10)$$

where

$$A_n(\beta) \equiv \int_0^\infty (1 + \beta u)^{n-2} e^{-u} du. \quad (3.11)$$

Appendix 3.8 provides a detailed development leading to Eqs. (3.9)-(3.11).  $A_n$  can also be rewritten in terms of the upper incomplete gamma function, but the present integral representation will be more useful later on.



Although it is not possible to write  $\beta_c$  in closed form, we provide bounds for  $A_n(\beta)$  :

$$1 \leq A_n(\beta) \leq \frac{1}{1 - \beta(n_{\max} - 2)} .$$

Details of this result are presented in Appendix 3.9. These inequalities lead to lower and upper bounds on the epidemic threshold

$$\beta_c \geq \frac{1}{\Omega(g_m, p_n) + (n_{\max} - 2)} , \quad (3.12a)$$

$$\beta_c \leq \frac{1}{\Omega(g_m, p_n)} , \quad (3.12b)$$

where we defined the *coupling* between groups as

$$\Omega(g_m, p_n) \equiv \left( \frac{\langle m(m-1) \rangle}{\langle m \rangle} \right) \left( \frac{\langle n(n-1) \rangle}{\langle n \rangle} \right) , \quad (3.13)$$

the product of the average excess group size and the average excess membership. If we take a random node within a group,  $\Omega(g_m, p_n)$  corresponds to its average number of *external neighbors*. It is therefore a good measure of the interaction of groups with one another.

### 3.5.2 Behavior for heterogeneous membership and group size

Let us consider power-law distributions  $p_n \propto n^{-\gamma_n}$  and  $g_m \propto m^{-\gamma_m}$  with large cutoffs  $n_{\max} \gg 1$  and  $m_{\max} \gg 1$ . We set  $\gamma_n, \gamma_m > 2$  so that  $\langle n \rangle$  and  $\langle m \rangle$  remain bounded.

For reasons that will become clear in Sec. 3.6, we distinguish a *strong* group coupling [ $\Omega(g_m, p_n) \gg n_{\max}$ ] from a *weak* group coupling [ $\Omega(g_m, p_n) \ll n_{\max}$ ]. Figure 3.3 illustrates the stationary state properties of the dynamics for two different pairs of exponents  $(\gamma_m, \gamma_n)$ , Fig. 3.3(a-c) corresponding to a strong group coupling and Fig. 3.3(d-f) corresponding to a weak group coupling.

Comparing Fig. 3.3(a) and Fig. 3.3(d), we note that all  $s_m$  decrease faster in the former case as the ratio  $\beta/\beta_c$  increases. From Eq. (3.5a), this is explained by a faster increase of the mean field  $r$ , resulting directly from a stronger coupling between groups.

The difference between Fig. 3.3(b) and 3.3(e) is more striking. While the group prevalence  $I_n$  does not vary much with  $n$  in Fig. 3.3(b)—the coupling  $\Omega(g_m, p_n)$  is strong—we observe a sequential activation of the groups for the weakly coupled system in Fig. 3.3(e). Figures 3.3(c) and 3.3(f) provide an even clearer illustration for a fixed  $\beta$ . When the coupling is strong, all distributions  $c_{n,i}$  are concentrated around roughly the same fraction of infected nodes within the groups. Weak coupling yields a more diverse scenario where smaller groups have very few infected nodes while the prevalence in larger groups can be very high. We qualify the latter as *active* groups. For groups of moderate size (e.g.,  $n = 50$ ),  $c_{n,i}$  is bimodal and highly dispersed, akin to a system near a critical point.

This is a telling illustration of why stochastic dynamics on networks with a high level of group organization are best described by approximate master equations : groups of nodes can have heterogeneous state distributions, and a cruder approximation (e.g., models averaging  $i/n$  for all groups of a given size or other mean-field approximations) is likely to miss many rich features of the dynamics. These features may be interesting by themselves, and important for the overall evolution of the process. While mean-field approaches are sometime *qualitatively* correct, they are most often *quantitatively* off

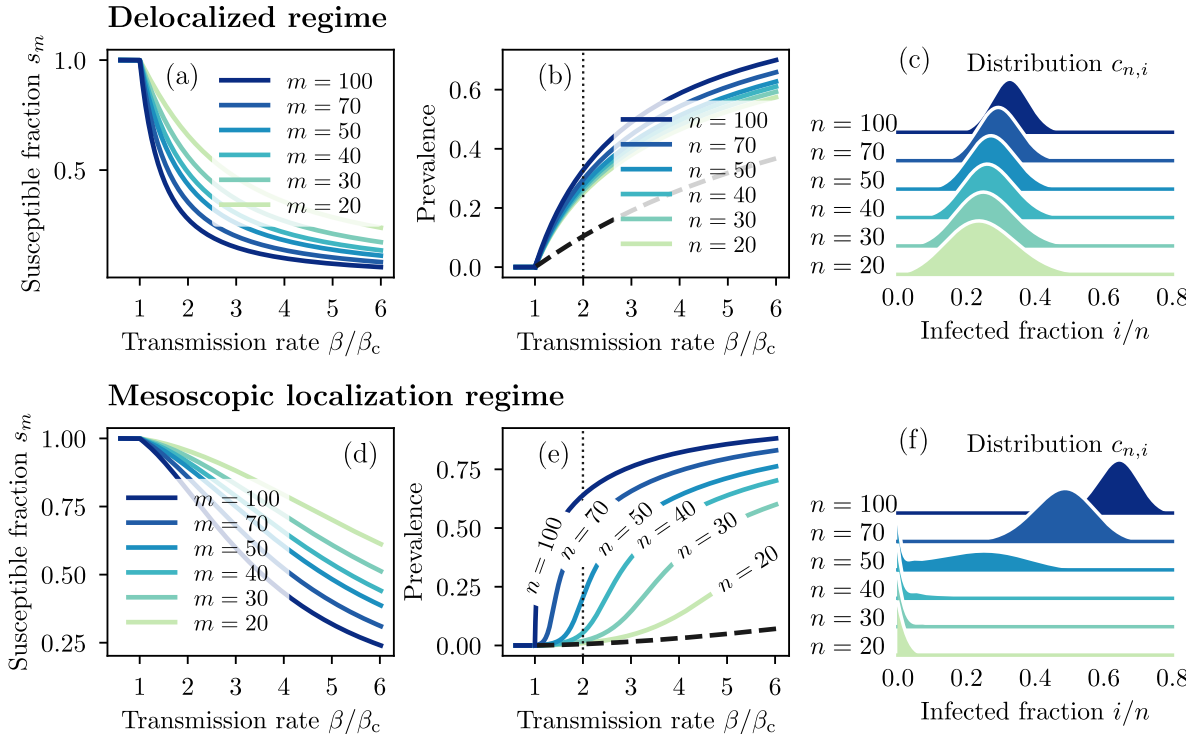


FIGURE 3.3 – Comparison of the stationary state near the epidemic threshold in the delocalized and mesoscopic localization regimes. Stationary state solutions were obtained from Eqs. (3.5a) and (3.6) for heterogeneous membership and group size distributions of the form  $g_m \propto m^{-\gamma_m}$  and  $p_n \propto n^{-\gamma_n}$ . We used  $m, n \in \{2, \dots, 100\}$ . The epidemic threshold  $\beta_c$  is the solution to Eq. (3.10). (a) and (d) Stationary fraction of susceptible nodes with a given membership  $m$  as a function of the transmission rate. (b) and (e) Group prevalence  $I_n$  (solid lines) and global prevalence  $I$  (dashed line) as a function of the transmission rate. (c) and (f) Distributions for the number of infected nodes  $i$  in a group of size  $n$ , obtained for  $\beta = 2\beta_c$  corresponding to the vertical dotted lines in the panel on their left. Spline interpolations are used for visual purpose. Upper row (a)-(c)  $\gamma_m = \gamma_n = 2.2$ . Lower row (d)-(f)  $\gamma_m = 4$  and  $\gamma_n = 3.5$ .

the mark [105, 106]. Approximate master equations yield both qualitatively and quantitatively correct results (see Appendix 3.11), ensuring that the observed phenomena are true properties of the original stochastic process.

The scenario presented by Figs. 3.3(e) and 3.3(f) is typical of a *smearred phase transition*. Instead of a clean critical point driven by a collective ordering, subparts of the system self-activate independently from the rest, as shown by the local order parameters  $I_n$ . This behavior has an intuitive explanation. Since  $p_n \propto n^{-\gamma_n}$ , a small proportion of the groups are very large, albeit of finite size. Near  $\beta_c$ , the largest groups are able to self-sustain an endemic state by themselves, but since the coupling is weak, the contagion does not spread through the rest of the network. As  $\beta$  increases beyond  $\beta_c$ , more groups are able to self-sustain a local outbreak, until a point where the epidemic delocalizes and invades the whole network. This analytical description is in line with the work of Ref. [64], where numerical evidence for Griffiths phases was found in a similar setting.

To predict the emergence of this phenomenon, we need to have some better intuition of the behavior

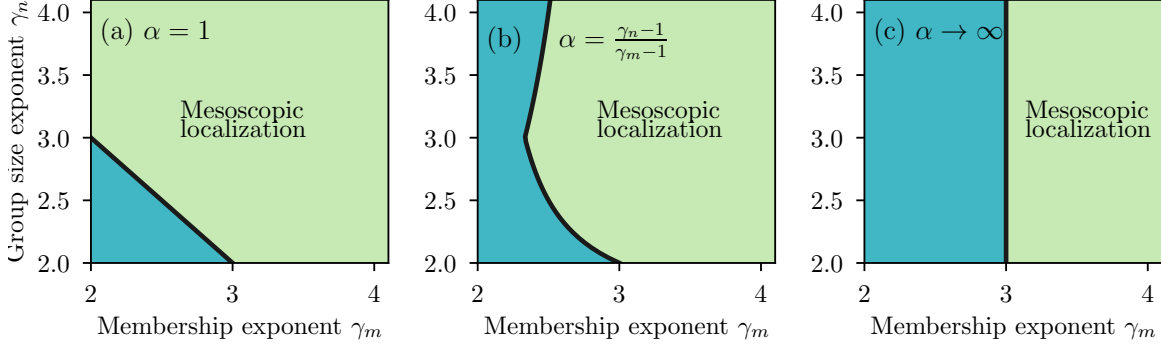


FIGURE 3.4 – Asymptotic localization regimes for power-law membership and group size distributions, and for different cutoff relationships  $m_{\max} \sim n_{\max}^\alpha$ . In the pale green regions, the epidemic is localized near the epidemic threshold  $\beta_c$ , while it is delocalized in the darker blue regions. The boundary separating the two regions is inferred from Eqs (3.20).

of  $I_n$  near  $\beta_c$ . Since  $\rho \rightarrow 0$  near the critical point, we write

$$I_n = \frac{1}{n} H'_n(1; \beta) \rho + O(\rho^2).$$

Performing a saddle-point approximation for  $H'_n(1; \beta)$ , we obtain the following asymptotic behavior for large  $n$

$$H'_n(1; \beta) \sim \begin{cases} \frac{n}{1 - \beta n} & \text{if } \beta < n^{-1} \\ n^{3/2} (\beta n)^n e^{-n+1/\beta} & \text{if } \beta \geq n^{-1}, \end{cases} \quad (3.14)$$

where “ $\sim$ ” means *asymptotically proportional*. For  $\beta = an^{-1}$  where  $a > 1$  is a constant independent of  $n$ , this implies that  $I_n \sim n^{1/2} e^{bn}$  with  $b > 0$ . A more formal proof could be made following an argument similar to the one used in Appendix 3.9. Therefore, near the epidemic threshold (i.e.,  $\beta = \beta_c + \epsilon$  with  $\epsilon \ll 1$ ), we expect the epidemic to be *localized* within any groups of size  $n$  for which  $\beta > n^{-1}$ . More formally, we say that the epidemic is localized near the epidemic threshold when  $I_{n_{\max}}/I_2 \sim n_{\max}^{1/2} e^{bn_{\max}}$ , and we then expect a smeared phase transition, such as the one presented in Fig. 3.3(e). Conversely, if  $I_{n_{\max}}/I_2 = O(1)$  near  $\beta_c$ , then we say that the epidemic is *delocalized*, and we expect a phase transition similar to the one shown in Fig. 3.3(b).

## 3.6 Mesoscopic localization

In this section, we fully characterize the emergence of *mesoscopic localization*, where the epidemic is localized only within the largest groups near  $\beta_c$  for power-law distributions of membership and group size. In Sec. 3.6.1, we derive general asymptotic expressions to distinguish the localization regimes, establishing a partition of the  $(\gamma_m, \gamma_n)$  space. We then investigate in Sec. 3.6.2 the effect of finite cutoffs on the localization regimes, and how our results relate to earlier works using the inverse participation ratio.

### 3.6.1 Asymptotic localization regimes

Let us assume that both  $n_{\max} \rightarrow \infty$  and  $m_{\max} \rightarrow \infty$ . As it will be shown, the relation between the cutoffs  $n_{\max}$  and  $m_{\max}$  influences the localization regimes. Henceforth, let us assume a general

asymptotic relationship of the form

$$m_{\max} \sim n_{\max}^{\alpha}, \quad (3.15)$$

where the exponent  $\alpha \geq 0$  encodes how both limits  $n_{\max} \rightarrow \infty$  and  $m_{\max} \rightarrow \infty$  are taken.

To gain some insights on the meaning of Eq. (3.15), let us assume for the sake of the argument that we have a finite-size network with  $N$  nodes and  $\langle m \rangle N / \langle n \rangle \sim N$  groups. We could impose cutoffs that are agnostic to the underlying distribution  $g_m$  and  $p_n$ , for instance  $m_{\max} \sim N^{1/2}$  and  $n_{\max} \sim N^{1/2}$ . This would correspond to  $\alpha = 1$ . Another option, borrowed from extreme value theory, would be to use the *natural* cutoffs of the two power-law distributions,  $m_{\max} \sim N^{1/(\gamma_m - 1)}$  and  $n_{\max} \sim N^{1/(\gamma_n - 1)}$  [38, 47]. This would correspond to  $\alpha = (\gamma_n - 1)/(\gamma_m - 1)$ . Finally, fixing one of the two cutoffs while letting the other go to infinity would correspond to the limit cases  $\alpha \rightarrow 0$  or  $\infty$ .

We now turn to the extraction of the asymptotic behavior of the epidemic threshold in the limit  $n_{\max} \rightarrow \infty$  for different combinations of  $\gamma_n$  and  $\gamma_m$ —this will inform us on the type of phase transition, i.e., a localized or a delocalized one.

First, we obtain a tighter upper-bound on  $\beta_c$  in the limit  $n_{\max} \rightarrow \infty$  for power-law group size distributions  $p_n \propto n^{-\gamma_n}$ . Formally, there exists some  $n' \in \mathbb{N}$  such that for all  $n_{\max} > n'$ ,

$$\beta_c \leq \min \left[ \frac{1}{\Omega(g_m, p_n)}, \frac{1}{n_{\max} - 2} \right]. \quad (3.16)$$

Details are provided in Appendix 3.9, but the general idea is to combine Eq. (3.12b) with another bound found by forbidding  $A_{n_{\max}}$  to grow exponentially with  $n_{\max}$ . The lower bound of Eq. (3.12a) and the upper bound of Eq. (3.16) tightly constrain the asymptotic behavior of  $\beta_c$ , which we write as

$$\beta_c^{-1} \sim \Omega(g_m, p_n) + n_{\max}. \quad (3.17)$$

Second, let us examine the asymptotic behavior of the coupling  $\Omega(g_m, p_n)$ . The first factor in Eq. (3.13) has the following behavior

$$\frac{\langle m(m-1) \rangle}{\langle m \rangle} \sim \begin{cases} n_{\max}^{\alpha(3-\gamma_m)} & \text{if } \gamma_m < 3, \\ \alpha \ln n_{\max} & \text{if } \gamma_m = 3, \\ 1 & \text{if } \gamma_m > 3, \end{cases} \quad (3.18)$$

and the second one has a similar form

$$\frac{\langle n(n-1) \rangle}{\langle n \rangle} \sim \begin{cases} n_{\max}^{3-\gamma_n} & \text{if } \gamma_n < 3, \\ \ln n_{\max} & \text{if } \gamma_n = 3, \\ 1 & \text{if } \gamma_n > 3. \end{cases} \quad (3.19)$$

Combining Eqs. (3.18) and (3.19) for different  $\gamma_m$  and  $\gamma_n$  leads to different scalings for  $\Omega(g_m, p_n)$ .

As a result, we find three cases for the scaling of  $\beta_c$  in the limit  $n_{\max} \rightarrow \infty$ :

1.  $\Omega(g_m, p_n) n_{\max}^{-1} \rightarrow \infty \Rightarrow \beta_c n_{\max} \rightarrow 0$ ,
2.  $\Omega(g_m, p_n) n_{\max}^{-1} \rightarrow O(1) \Rightarrow \beta_c n_{\max} \rightarrow q < 1$ ,
3.  $\Omega(g_m, p_n) n_{\max}^{-1} \rightarrow 0 \Rightarrow \beta_c n_{\max} \rightarrow 1$ .

This classification allows us to associate an asymptotic behavior to each pair  $(\gamma_m, \gamma_n)$ . If  $\gamma_m \geq 3$ , we necessarily have  $\beta_c n_{\max} \rightarrow 1$ . Otherwise, if  $2 < \gamma_m < 3$  and

- $2 < \gamma_n < 3$ , then

$$\beta_c n_{\max} \rightarrow \begin{cases} 0 & \text{if } 3 - \gamma_n + \alpha(3 - \gamma_m) > 1, \\ q < 1 & \text{if } 3 - \gamma_n + \alpha(3 - \gamma_m) = 1, \\ 1 & \text{if } 3 - \gamma_n + \alpha(3 - \gamma_m) < 1, \end{cases} \quad (3.20a)$$

- $\gamma_n = 3$ , then

$$\beta_c n_{\max} \rightarrow \begin{cases} 0 & \text{if } \alpha(3 - \gamma_m) \geq 1, \\ 1 & \text{if } \alpha(3 - \gamma_m) < 1, \end{cases} \quad (3.20b)$$

- $\gamma_n > 3$ , then

$$\beta_c n_{\max} \rightarrow \begin{cases} 0 & \text{if } \alpha(3 - \gamma_m) > 1, \\ q < 1 & \text{if } \alpha(3 - \gamma_m) = 1, \\ 1 & \text{if } \alpha(3 - \gamma_m) < 1. \end{cases} \quad (3.20c)$$

Note that the asymptotic behavior  $\beta_c n_{\max} \rightarrow q < 1$  never fills an area in the  $(\gamma_m, \gamma_n)$  space—it is simply a limiting case. The two other cases fill the  $(\gamma_m, \gamma_n)$  space, and we interpret them as different *localization regimes* using the definitions of Sec. 3.5.2. In the region where  $\beta_c n_{\max} \rightarrow 0$ , we have  $I_{n_{\max}}/I_2 = O(1)$  near the epidemic threshold, and the epidemic is delocalized since groups of all sizes are involved. In the region where  $\beta_c n_{\max} \rightarrow 1$ , we have, near the epidemic threshold ( $\beta = \beta_c + \epsilon$  with  $\epsilon \ll 1$ ),  $\beta > n_{\max}^{-1}$ , and therefore  $I_{n_{\max}}/I_2 \sim n_{\max}^{1/2} e^{\beta n_{\max}}$ . The epidemic is therefore localized, thriving only in the largest groups.

Equations (3.20) are thus used to identify the region where we expect mesoscopic localization, as illustrated in Fig. 3.4 for different values of  $\alpha$ . One striking observation is the ubiquity of mesoscopic localization : for a large portion of the parameter space, we expect a disease to be localized around the largest groups. It is worth recalling that the average degree of a node is proportional to  $\langle n(n-1) \rangle$ , hence sparse networks correspond only to the upper portion ( $\gamma_n > 3$ ) of the phase diagrams in Fig. 3.4.

### 3.6.2 Finite cutoffs and mesoscopic inverse participation ratio

The results of Sec. 3.6.1 were obtained in the asymptotic limit  $n_{\max} \rightarrow \infty$ . However, cutoffs in real systems are always finite. A finite value for  $n_{\max}$  relaxes the conditions defined in Eqs. (3.20a-c). For a pair  $(\gamma_m, \gamma_n)$  in the asymptotically localized regime, it is possible to have either  $\beta_c \geq n_{\max}^{-1}$ . To stay coherent with our definition for a localized epidemic, we must have  $\beta_c \geq n_{\max}^{-1}$ . Therefore, the condition  $\beta_c \equiv n_{\max}^{-1}$  leads to the finite cutoffs boundary, given by solutions to Eq. (3.10) in terms of  $\gamma_m$  and  $\gamma_n$ . In Fig. 3.5, we illustrate the boundary separating the delocalized and localized regimes for increasing values of  $n_{\max}$ , slowly converging on the asymptotic conditions. The size of the mesoscopic localization region is smaller compared with that of the asymptotic limit, but it still fills most of the parameter space corresponding to sparse networks ( $\gamma_n > 3$ ).

Another consequence of finite cutoffs is to blur the line between a localized and a delocalized epidemic. Taking pairs  $(\gamma_m, \gamma_n)$  closer to the finite-size boundary, we show how this affects the group prevalence in Fig. 3.6(a) and Fig. 3.6(b), with  $\beta_c < n_{\max}^{-1}$  and  $\beta_c > n_{\max}^{-1}$  respectively. Near  $\beta_c$ , we still associate

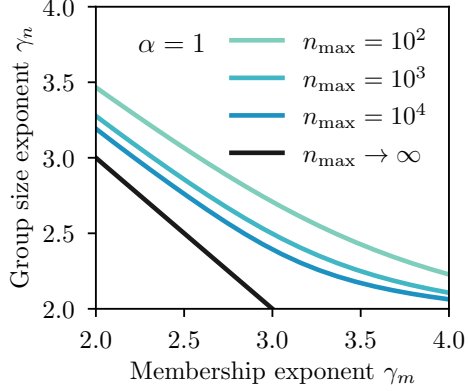


FIGURE 3.5 – Impact of finite cutoffs on the boundary separating the localized and delocalized regimes for power-law membership and group size distributions. We used  $m, n \in \{2, \dots, n_{\max}\}$ , hence  $\alpha = 1$ , and different values of  $n_{\max}$ . Finite cutoffs boundaries are obtained by imposing  $\beta_c = n_{\max}^{-1}$  and solving Eq. (3.10) for different pairs  $(\gamma_m, \gamma_n)$ . The asymptotic case  $n_{\max} \rightarrow \infty$  is obtained from Eqs. (3.20).

Fig. 3.6(a) and 3.6(b) with a delocalized and localized outbreak, respectively, but the difference is less marked compared to Fig. 3.3(a) and Fig. 3.3(b). Therefore, even though the dichotomy is sharp and clear in the asymptotic limit  $n_{\max} \rightarrow \infty$ , we need to keep in mind that for realistic systems, localization lives on a spectrum. Our next goal is to *quantify* mesoscopic localization.

*At the node level*, an epidemic is considered localized if the contagion is mostly present within a subset of nodes  $\mathcal{L} \subset \mathcal{V}$ , referred to as the localization set, and  $\mathcal{V} = \{1, \dots, N\}$  is the set of all nodes. An important result from quenched mean-field theory is that the marginal probability for each node  $j$  of being infected near  $\beta_c$  is proportional to  $v_j$ , where  $\{v_j\}_{j \in \mathcal{V}}$  are the elements of the principal eigenvector (PEV) of the adjacency matrix. Epidemic localization can thus be mapped onto eigenvector localization [46, 190, 229, 230, 252]. With a normalized eigenvector satisfying  $\sum_j v_j^2 \equiv 1$ , a completely delocalized epidemic at the level of nodes implies  $v_j \sim N^{-1/2} \forall j \in \mathcal{V}$ , while a purely localized one corresponds to  $v_j \sim |\mathcal{L}|^{-1/2} \forall j \in \mathcal{L}$  and  $v_j \sim 0 \forall j \notin \mathcal{L}$ . A standard scalar to quantify the localization is the inverse participation ratio  $Y_4(N)$ . We use the following rescaled version

$$\tilde{Y}_4(N) \equiv N \sum_{j=1}^N v_j^4. \quad (3.21)$$

For a delocalized eigenvector,  $\tilde{Y}_4(N) \sim 1$ , while for a localization set of size  $|\mathcal{L}| \sim N^\delta$  and  $\tilde{Y}_4(N) \sim N^{1-\delta}$ . Consequently,  $\tilde{Y}_4^{-1}$  is an effective measure for the fraction of nodes belonging to the localization set.

*At the mesoscopic level*, we consider an epidemic localized if the contagion is mostly present within a subset of the groups. The difference is subtle, but important : if we observe a delocalized epidemic at the mesoscopic level, it could still be localized at the level of nodes. To quantify mesoscopic localization, we use an inverse participation ratio as well

$$\hat{Y}_4(p_n) = \frac{\sum_n p_n I_n^4}{(\sum_n p_n I_n^2)^2} = \begin{cases} 1 & \text{if } I_n \propto 1 \forall n, \\ p_{n'}^{-1} & \text{if } I_n \propto \delta_{n,n'}. \end{cases} \quad (3.22)$$

As a result,  $\hat{Y}_4^{-1}$  is an effective measure for the fraction of *groups* participating in the epidemic. Interestingly, Eq. (3.22) can be obtained with our analytical formalism, using  $I_n$  evaluated at the

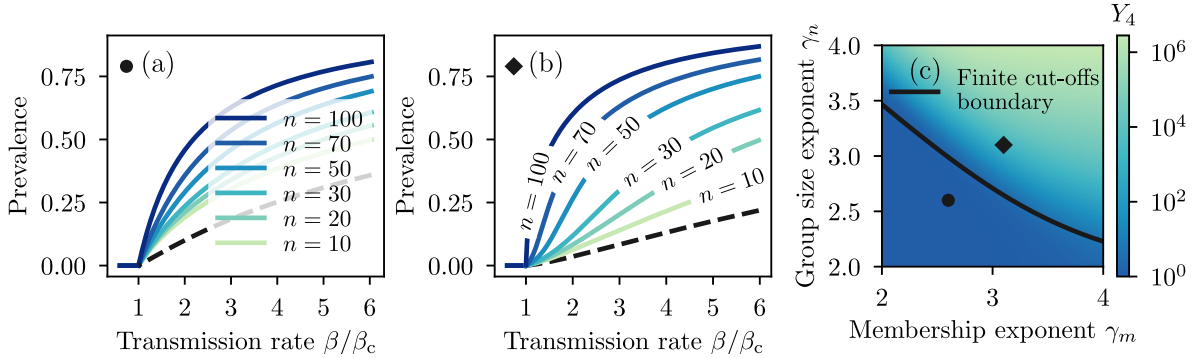


FIGURE 3.6 – Comparison of the stationary state and the level of localization near the epidemic threshold for networks closer to the finite cutoffs boundary. We used power-law membership and group size distributions with  $m, n \in \{2, \dots, 100\}$ . (a)-(b) Group prevalence (solid lines) and global prevalence (dashed line) as a function of the transmission rate. (a)  $\gamma_m = \gamma_n = 2.6$ , yielding  $\beta_c < n_{\max}^{-1}$ . (b)  $\gamma_m = \gamma_n = 3.1$ , yielding  $\beta_c > n_{\max}^{-1}$ . (c) Quantification of the mesoscopic localization phenomenon using the inverse participation ratio defined in Eq. (3.22). The solid line corresponds to the boundary between the localized and delocalized regimes, obtained by imposing  $\beta_c = n_{\max}^{-1}$  and solving Eq. (3.10) for different pairs  $(\gamma_m, \gamma_n)$ .

epidemic threshold  $\beta_c$ , or through the connection with quenched mean-field theory. In the latter case, one extracts the PEV of a network with cliques, then computes

$$I_n \propto \frac{1}{|\mathcal{C}_n|} \sum_{\mathcal{S} \in \mathcal{C}_n} \sum_{j \in \mathcal{S}} \frac{v_j}{n}, \quad (3.23)$$

where  $\mathcal{C}_n$  is the set of cliques of size  $n$  and  $\mathcal{S}$  is the set of nodes belonging to a specific clique. The group distribution then correspond to  $p_n \propto |\mathcal{C}_n|$ . Note that this measure relies on an explicit knowledge of  $\mathcal{C}_n$ , which is already given for synthetic networks (see Appendix 3.10), or could be extracted using a clique decomposition for real networks.

In Fig. 3.6(c), we illustrate the behavior of  $\widehat{Y}_4(p_n)$  as a function of  $(\gamma_m, \gamma_n)$ , obtained with our analytical formalism. As expected, the inverse participation ratio changes drastically near the boundary separating the delocalized and localized regimes for finite cutoffs. The change would become sharper and sharper as we let  $n_{\max} \rightarrow \infty$ , and the position of the boundary would move closer to the asymptotic limit, as in Fig. 3.5. This inverse participation ratio is therefore a good measure for mesoscopic localization, and could be used to get insights on how the epidemic changes from a localized to a delocalized phase as we increase  $\beta$  beyond  $\beta_c$ .

In Fig. 3.7, we compare the finite-size scaling of the inverse participation ratios for nodes and groups, obtained by generating synthetic networks in the delocalized and localized regime and extracting their PEV. Although our analytical formalism effectively describes groups of a subextensive size, this is not a necessary condition to observe mesoscopic localization. We have therefore relaxed this assumption to generate the synthetic networks : we have used cutoffs that scale with the number of nodes  $m_{\max} = N^{1/(\gamma_m - 1)}$  and  $n_{\max} = N^{1/(\gamma_n - 1)}$ . These are more appropriate for the finite-size scaling analysis.

In Fig. 3.7(a), we see that the inverse participation ratio for nodes  $\widetilde{Y}_4$  increases in both the delocalized and the mesoscopic localization regime. It scales similarly to the inverse fraction of the nodes belonging to the maximal  $K$  core, in agreement with previous works on the subject [45, 46, 230]. The localization

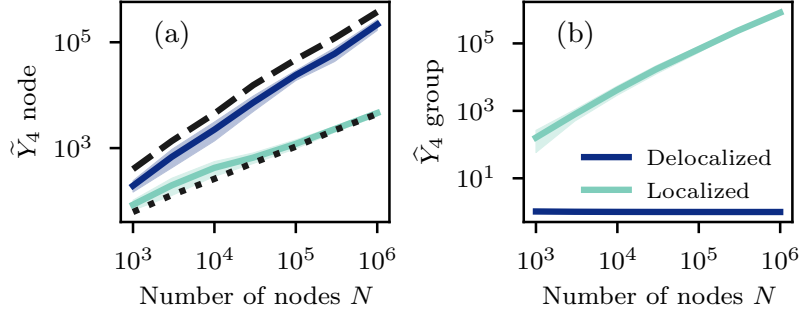


FIGURE 3.7 – Mesoscopic localization is imperceptible using the standard inverse participation ratio on nodes. We performed a finite-size scaling of the inverse participation ratio in the delocalized and mesoscopic localization regime, for nodes and groups. We generated multigraphs of various sizes with different power-law membership and group size distributions [see Appendix 3.10]. In the delocalized regime, we used  $\gamma_m = 2.3$  and  $\gamma_n = 3.5$ ; in the localized regime, we used  $\gamma_m = \gamma_n = 3.5$ . In both regimes,  $m \in \{2, \dots, m_{\max}\}$  and  $n \in \{2, \dots, n_{\max}\}$  with natural cutoffs  $m_{\max} = N^{1/(\gamma_m-1)}$  and  $n_{\max} = N^{1/(\gamma_n-1)}$ . (a) The solid lines represent the average inverse participation ratio  $\hat{Y}_4$  for nodes [Eq. (3.21)]. The dashed (dotted) line is the average inverse of the fraction of nodes associated to the maximal  $K$  core in the delocalized (localized) regime. (b) The solid lines represent the average inverse participation ratio  $\hat{Y}_4$  for groups [Eq. (3.22)]. We extracted  $I_n$  from the PEV. The shaded regions in both panels correspond to one standard deviation.

set can thus be associated with the innermost core in both cases, and despite a different scaling law there is no clear sign of a change of regime between the two curves. Figure 3.7(b) tells us another story : the inverse participation ratio for groups  $\hat{Y}_4$  converges to 1 in the delocalized regime, but scales as a power law in the mesoscopic localization regime, clearly indicating a transition of regime.

Figure 3.7 strongly advocates for a change of perspective if we want to detect the potentially hidden localized phase at the mesoscopic level. We need to focus on the higher-level organization, the groups, and find better ways to characterize their impact on the dynamics. If we focus our attention at the node level, Fig. 3.7(a) tells us that an epidemic localized at the mesoscopic level is no different from a delocalized one—the contagion is mostly present within the innermost core in both cases. However, the composition of this core and of the outer shells is quite different, as can be inferred from  $\hat{Y}_4$  in Fig. 3.7(b). In the localized regime, the innermost core is composed mostly of the largest groups, while groups of all sizes compose the core in the delocalized regime. Recall that a bias toward larger groups has dramatic consequences on the dynamics, leading to a smeared phase transition instead of a clean one.

Before closing this section, it is probably useful to stress once more the versatility and generality of our approach. The results on synthetic networks are representative of results that can also be obtained on real complex systems. More complex networks are generally made of mesoscopic substructures, dense or sparse—not necessarily cliques—with possibly weighted and directed edges. The important observation is that the details of these substructures do not matter much. As long as it is possible to identify them, using community detection [93, 94], random clique cover [294], or hypergraph reconstruction [300], one can construct a higher-order representation of the original network with nodes belonging to groups and evaluate the localization on these groups using a measure similar to Eq. (3.23). Following this line, the original structure does not even need to be a network of pairwise interactions. It could already be a higher-order representation, such as a simplicial complex or a hypergraph [22].



### 3.7 Discussion

One of the important factors behind the success of network science to study contagions, from infectious diseases to the spread of information, is that it provides a mathematical framework to go beyond the assumption of a homogeneous population [231]. Contagions are rarely driven by the average individual, mostly because some individuals are simply more connected than others but also potentially more central. Beyond the fact that they drive the dynamics of contagions, these key actors are also critical to their control. On the one hand, it allows the mathematical formulation of targeted immunization and interventions [139, 234] : which individuals should be immunized or removed from the network to minimize the spread on an infectious disease? On the other hand, it also permits the identification of influential spreaders [211] : which individuals should seed a contagion in order to maximize its spread? These different ideas all revolve around a control theory for contagions, but also all depend on a theoretical understanding of what type of structures matters for contagions.

In practice, however, social networks are not randomly mixed but contain a higher-level organization determined by workplaces, schools, events, etc. ; such that key actors can be places, social gatherings, or more abstractly groups, rather than the individuals themselves. Thankfully, multiple new approaches to handle higher-order interactions have been proposed in recent years. In the thermodynamic limit, the networks used in this chapter can equivalently be represented using ideas of topological simplexes from topology [142], hypergraphs [88, 147], or projections of bipartite networks [125, 218]. Under the right level of mean-field approximation, these are all equivalent. However, their dynamics at the mesoscopic level can be very heterogeneous, as in Fig. 3.3(f), since groups can take considerably more different states than individuals who are usually only susceptible or infected. Therefore, adequate care should be exercised not to over-simplify (coarse grain) the mathematical description in order to embrace this heterogeneity. Our group-based approximate master equation framework acknowledges fully this warning.

Using this approach, we have observed and analyzed a phenomenon of mesoscopic localization where contagions can concentrate around groups that are large enough to allow a local, self-sustained outbreak in the presence of some weak external group coupling. Interestingly, while there is little empirical evidence for localization of real contagions around hubs in a contact network, there are well-known cases of dynamics resembling mesoscopic localization. For example, bacterial infections in hospitals (e.g., *C. difficile* [197]) are already a well-documented example of mesoscopically localized contagions, but are simply never studied analytically as such.

In this mesoscopic localization phase, influential groups are naturally found to be the larger ones around which a contagion can localize. Intervention or control operating at a structural level (i.e., on groups rather than on individuals) should therefore focus around these influential groups. The large toolbox developed for targeted immunization [120] and identification of influential spreaders [293] could now be leveraged, at the mesoscopic level, to better understand and control contagions on networks capable of mesoscopic localization. In chapter 4, we investigate the impact of removing groups as a model of school closures and event cancellations. We find that delocalized dynamics are characterized by a linear relationship between outbreak size and the strength of our intervention, akin to mass-action models. Conversely, localized dynamics show a nonlinear relationship that varies with the importance of the localization effects. For strongly localized epidemics, there is an increasing effectiveness of interventions, leading to a sudden collapse of the epidemic.

In a broader context, higher-order structures were found to be important for a wide range of dynamics, from competitive dynamics [113] to social contagion [142]. Several of these studies highlight nontrivial effects of higher-order structures on dynamics using numerical tools or very coarse-grained analytical methods. These approaches, ignoring the heterogeneous states of groups, limit the type of questions and behaviors that can be answered and analyzed. We wish to emphasize that master equation descriptions provide valuable insights into the mechanisms of these dynamics and their interplay with higher-order structures. For instance, we conjecture that mesoscopic localization is even more present in systems with social reinforcement mechanisms [226], and that its impacts on the global state of the dynamics are even more dramatic.

There are now several avenues open to broaden the applicability of our simple approach. In its current form, the only inputs required are a membership distribution  $g_m$  and a group size distribution  $p_n$ , along with the specification of the local dynamics. As a first step, our future works will focus on improving the heterogeneous mean-field coupling between master equations. We could, for example, refine our description of the states of the nodes in order to capture dynamical correlations with the state of the groups in which they are found, include structural correlations between the memberships of nodes and the sizes of groups through a joint distributions  $P(m, n)$ , or allow groups with more complex inner contact patterns. We hope that our work on mesoscopic localization and the framework that has emerged will provide a solid foundation for the continuing efforts to improve our understanding of dynamics on complex networks.

### 3.8 Appendix : Detailed development for the epidemic threshold

The function  $F(\rho)$  corresponds to

$$F(\rho) = r(\rho) \left[ \frac{\sum_m m(m-1)s_m(\rho)g_m}{\sum_m ms_m(\rho)g_m} \right].$$

To find its derivative with respect to  $\rho$ , let us note that  $s_m(\rho) = 1 + O(\rho)$  and  $r(\rho) = O(\rho)$  as  $\rho \rightarrow 0$ , which can be deduced from Eq. (3.2a) by using  $c_{n,i} = \delta_{i,0} + h_{n,i}\rho + O(\rho^2)$ . Therefore,

$$\left. \frac{dF}{d\rho} \right|_{\rho \rightarrow 0} = \frac{\langle m(m-1) \rangle}{\langle m \rangle} \left. \frac{dr}{d\rho} \right|_{\rho \rightarrow 0}, \quad (3.24)$$

and the derivative of  $r(\rho)$  is

$$\left. \frac{dr}{d\rho} \right|_{\rho \rightarrow 0} = \frac{\sum_{n,i} \beta i(n-i)h_{n,i}p_n}{\sum_n np_n}. \quad (3.25)$$

In terms of the generating functions  $H_n(x; \beta)$ , with

$$\begin{aligned} H'_n(1; \beta) &= \sum_i ih_{n,i}, \\ H''_n(1; \beta) &= \sum_i i(i-1)h_{n,i}, \end{aligned}$$

we obtain the relation,

$$\sum_i i(n-i)h_{n,i} = (n-1)H'_n(1; \beta) - H''_n(1; \beta). \quad (3.26)$$

Combining Eqs. (3.24), (3.25), and (3.26), we arrive at

$$\left. \frac{dF}{d\rho} \right|_{\rho \rightarrow 0} = \beta \frac{\langle m(m-1) \rangle}{\langle m \rangle \langle n \rangle} \langle (n-1)H'_n(1; \beta) - H''_n(1; \beta) \rangle . \quad (3.27)$$

With the integral representation of  $H_n(x; \beta)$ , Eq. (3.8), and the derivatives

$$\begin{aligned} H'_n(1; \beta) &= n \int_0^\infty (1 + \beta u)^{n-1} e^{-u} du , \\ H''_n(1; \beta) &= n(n-1) \int_0^\infty (1 + \beta u)^{n-2} \beta u e^{-u} du , \end{aligned}$$

we end up with the simplification

$$(n-1)H'_n(1; \beta) - H''_n(1; \beta) = n(n-1)A_n(\beta) ,$$

where

$$A_n(\beta) \equiv \int_0^\infty (1 + \beta u)^{n-2} e^{-u} du . \quad (3.28)$$

Inserting in Eq. (3.27) and setting the derivative to 1, we finally obtain the implicit expression for the epidemic threshold given by Eq. (3.10).

### 3.9 Appendix : Bounds on the epidemic threshold

Let us bound  $\beta_c$  for any  $n_{\max}$  by bounding  $A_n(\beta)$  [Eq. (3.28)] for all  $n$ . First, since  $\beta u \geq 0$ , then

$$A_n(\beta) \geq \int_0^\infty e^{-u} du = 1 . \quad (3.29)$$

Second, we rewrite

$$A_n(\beta) = \int_0^\infty e^{\phi(u; \beta)} du ,$$

where  $\phi(u; \beta) = (n-2) \ln(1 + \beta u) - u$ . Since  $\ln(1 + x) \leq x$ ,  $\phi(u; \beta) \leq \beta(n-2)u - u$ , which implies

$$A_n(\beta) \leq \begin{cases} \infty & \text{if } \beta(n-2) \geq 1 , \\ \frac{1}{1 - \beta(n-2)} & \text{if } \beta(n-2) < 1 . \end{cases}$$

We relax the conditions by replacing  $n$  by  $n_{\max}$  everywhere on the right-hand side

$$A_n(\beta) \leq \begin{cases} \infty & \text{if } \beta(n_{\max} - 2) \geq 1 , \\ \frac{1}{1 - \beta(n_{\max} - 2)} & \text{if } \beta(n_{\max} - 2) < 1 . \end{cases} \quad (3.30)$$

By inserting Eqs. (3.29) and (3.30) in Eq. (3.10) and solving for  $\beta$ , we find the bounds of Eq. (3.12). Only the second case of Eq. (3.30) leads to a coherent bound for  $\beta_c$ .

The upper bound on the epidemic threshold is not very tight, but we can do better if we assume  $p_n \propto n^{-\gamma_n}$  and the limit  $n_{\max} \rightarrow \infty$ . It follows that there exists some  $n' \in \mathbb{N}$  such that for all  $n_{\max} > n'$  the epidemic threshold must respect  $\beta_c \leq (n_{\max} - 2)^{-1}$ . Let us make a proof by contradiction : we

start with the premise that  $\beta = a(n_{\max} - 2)^{-1}$  for some arbitrary constant  $a > 1$ . We know that  $\ln(1+x) \geq x(1-x)$  for all  $x \geq 0$ , hence

$$\phi(u; \beta) \geq -\beta^2(n-2)u^2 + [\beta(n-2) - 1]u .$$

Making the change of variable  $y = \beta\sqrt{n-2}u$  and defining  $d \equiv [\beta(n-2) - 1]/(2\beta\sqrt{n-2})$ , we arrive at

$$\begin{aligned} A_n(\beta) &\geq \frac{e^{d^2}}{\beta\sqrt{n-2}} \int_0^\infty e^{-(y-d)^2} dy \quad \forall n > 2 \\ &\geq \frac{\sqrt{\pi}e^{d^2}}{2\beta\sqrt{n-2}} . \end{aligned}$$

Let us focus on  $n = n_{\max}$ . In this case, using our premise for  $\beta$ , we have

$$d^2 = \frac{(a-1)^2(n_{\max}-2)}{4a^2} \equiv b(n_{\max}-2) ,$$

where  $b > 0$ . Therefore, there always exists a constant  $B_1 > 0$  independent from  $n_{\max}$  and  $a$  such that

$$A_{n_{\max}}(\beta) \geq \frac{1}{a} \left( \frac{1}{2} \sqrt{\pi(n_{\max}-2)} e^{-2b} \right) e^{bn_{\max}} \geq \frac{B_1}{a} e^{bn_{\max}} .$$

This provides a lower bound for the following term :

$$\begin{aligned} \langle n(n-1)A_n(\beta) \rangle &\geq p_{n_{\max}} n_{\max} (n_{\max}-1) A_{n_{\max}}(\beta) \\ &\geq \frac{B_2}{a} n_{\max}^{2-\gamma_n} e^{bn_{\max}} , \end{aligned}$$

where we assumed  $p_n \propto n^{-\gamma_n}$  with  $\gamma_n < \infty$ . For some constant  $B_2 < \infty$ . Inserting this and our premise on  $\beta$  in Eq. (3.10), we obtain an expression of the form

$$n_{\max}^{1-\gamma_n} \leq B_3 e^{-bn_{\max}} . \quad (3.31)$$

For some constant  $B_3 < \infty$ . Equation (3.31) is clearly not respected in the limit  $n_{\max} \rightarrow \infty$ , hence completing the proof by contradiction.

Note that a solution  $\beta_c > (n_{\max}-2)^{-1}$  is not ruled out if  $p_n$  decrease exponentially for large  $n$ .

### 3.10 Appendix : Generation of networks with cliques

We generated multigraphs using a stub matching process. First, each node  $j \in \mathcal{V}$  is assigned a membership  $m$  drawn from  $g_m$ , resulting in a membership sequence  $\mathbf{m} = [m_1, m_2, \dots, m_N]$ . Then, we create a group size sequence of length  $N'$ ,  $\mathbf{n} = [n_1, n_2, \dots, n_{N'}]$ , by drawing sizes  $n_k$  according to  $p_n$ . We must additionally constrain the sequence such that the number of membership stubs and the number of group stubs (available spot for the nodes) are the same

$$\sum_{j=1}^N m_j = \sum_{k=1}^{N'} n_k . \quad (3.32)$$

In practice, if the right-hand side of Eq. (3.32) is smaller than the left-hand side, we add another group with size  $n$  drawn from  $p_n$ . If it is bigger, we remove a group uniformly at random. We repeat this process until the number of stubs is equal on both sides.  $N'$  is therefore not fixed, but it is expected that  $N' \simeq \langle m \rangle N / \langle n \rangle$  for large  $N$  since both  $\langle m \rangle$  and  $\langle n \rangle$  are bounded.

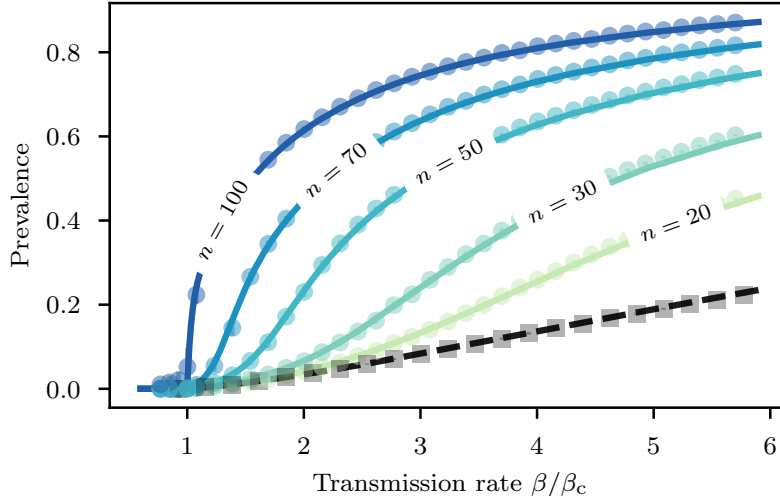


FIGURE 3.8 – Validation of the approximate master equation approach with Monte Carlo simulations, using the efficient algorithm developed in Appendix A. We used a homogeneous membership distribution  $g_m = \delta_{m,3}$ , a heterogeneous group size distribution of the form  $p_n \propto n^{-\gamma_n}$ , and  $n \in \{2, \dots, 100\}$ . The solid lines represent the group prevalence and the dashed line represents the global prevalence obtained with Eqs. (3.5a) and (3.6). The markers represent the average group (circles) and global (squares) prevalences in the quasistationary state over 20 network realizations of size  $N = 5 \times 10^5$ . The error bars (smaller than the markers) represent the standard deviation over these 20 realizations. For each network, 20% of the nodes were initially infected at random, then after a burn-in period  $\Delta t \in [50, 1000]$ , between 100 and 1000 states have been sampled to estimate the prevalences, each separated by a decorrelation period  $\Delta t \in [1, 10]$ . Larger values for the burn-in period, the decorrelation period and the number of states sampled were used near the epidemic threshold. The epidemic threshold  $\beta_c$  is the solution to Eq. (3.10).

Once we have the membership and group size sequences, we match the stubs uniformly at random—an edge is added between each pair of nodes belonging to the same group. This effectively creates loopy multigraphs, but the loops and multi-edges represent a vanishing fraction of the total number of edges for  $N \rightarrow \infty$ ; we do not remove them since they have a marginal impact on the dynamics.

### 3.11 Appendix : Validation with Monte Carlo simulations

In Fig. 3.8, we compare the predictions of our approximate master equation approach with the results of extensive Monte Carlo simulations. Our analytical framework accurately reproduces the behavior of the SIS model on synthetic networks generated with the method presented in Appendix 3.10. Figure 3.8 also confirms the existence of the mesoscopic localization phenomenon predicted by our approach.

To simulate the SIS model on multigraphs, we used the efficient algorithm developed in Appendix A to evolve the state of the system. Since the system typically reaches the absorbing state for finite-size networks near the epidemic threshold, we sampled the *quasistationary state* using the state-of-the-art method presented in Refs. [68, 248]. We kept a history of 100 previous states that were each updated at a rate  $\omega \in [10^{-3}, 10^{-2}]$  by the current state of the system. If the system fell on the absorbing state, it was replaced by a random state picked in the history—after a sufficient burn-in period, this method samples the quasistationary state.

### 3.12 Appendix : Scaling the transmission rate with group size

In the approximate master equations (3.1), an infected node in a group transmits the disease to all susceptible nodes at rate  $\beta$ . Even though it is reasonable to have more infections within large groups, an individual might not interact with all others as much as within smaller groups. For instance, assume two groups of size  $n$  and  $n'$ , with  $n > n'$ , the first representing a workplace and the second representing a household. An infected individual belonging to both interacts with more people in the first, but the strength of the interaction is more important in the second.

Fortunately, our framework is highly flexible. We could replace the term  $\beta i$  in Eq. (3.1b) by a general infection function  $f(n, i)$  for the nodes in the group. For the matter at hand, we simply scale the transmission rate as  $\beta \mapsto \beta n^{-\nu}$  with  $0 \leq \nu \leq 1$ , assuming that, on average, the interaction strength decreases with the group size. The analysis already performed for  $\nu = 0$  is extended to arbitrary values of  $\nu$  by direct substitutions.

#### 3.12.1 Epidemic threshold

First, we have the following new definition for  $r(\rho)$  in the stationary state

$$r(\rho) = \frac{\sum_{n,i} \beta n^{-\nu} i (n-i) c_{n,i}(\rho) p_n}{\sum_{n,i} (n-i) c_{n,i}(\rho) p_n}.$$

Near the absorbing state, we redefine the generating function as

$$H_n(x; \beta, \nu) = h_{n,0} + \frac{n^\nu}{\beta} \int_0^\infty \left[ \left( 1 + \frac{\beta u x}{n^\nu} \right)^n - 1 \right] u^{-1} e^{-u} du.$$

The condition for the epidemic threshold then becomes

$$\begin{aligned} \left. \frac{dF}{d\rho} \right|_{\rho \rightarrow 0} &= \beta \frac{\langle m(m-1) \rangle}{\langle m \rangle \langle n \rangle} \\ &\quad \times \langle n^{-\nu} [(n-1)H'_n(1; \beta, \nu) - H''_n(1; \beta, \nu)] \rangle \\ &\equiv 1. \end{aligned}$$

After some algebraic manipulations, we arrive at a new implicit expression for  $\beta_c$ ,

$$\beta \frac{\langle m(m-1) \rangle}{\langle m \rangle \langle n \rangle} \langle n^{1-\nu} (n-1) A_n(\beta, \nu) \rangle = 1, \quad (3.33)$$

where

$$A_n(\beta, \nu) \equiv \int_0^\infty \left( 1 + \frac{\beta u}{n^\nu} \right)^{n-2} e^{-u} du.$$

Since  $A_n(\beta, \nu)$  has a similar form as Eq. (3.28), it is straightforward to reproduce the results of Appendix 3.9 in this more general context. For a power-law distribution  $p_n \sim n^{-\gamma n}$ , we have the following asymptotic behavior for the epidemic threshold

$$\beta_c^{-1} \sim \Omega(g_m, p_n; \nu) + n_{\max}^{1-\nu}, \quad (3.34)$$

where the *coupling* between groups is

$$\Omega(g_m, p_n; \nu) = \left( \frac{\langle m(m-1) \rangle}{\langle m \rangle} \right) \left( \frac{\langle n^{1-\nu} (n-1) \rangle}{\langle n \rangle} \right). \quad (3.35)$$

### 3.12.2 Behavior near the absorbing state

The group prevalence  $I_n$  near the absorbing state can be estimated from a saddle-point approximation of  $H'_n(1; \beta, \nu)$  as well. For large  $n$ ,

$$H'_n(1; \beta, \nu) \sim \begin{cases} \frac{n}{1 - \beta n^{1-\nu}} & \text{if } \beta < n^{\nu-1}, \\ n^{3/2} (\beta n^{1-\nu})^n e^{-n+n^\nu/\beta} & \text{if } \beta \geq n^{\nu-1}. \end{cases} \quad (3.36)$$

For  $\beta = an^{\nu-1}$  where  $a > 1$  is a constant independent of  $n$ , we still have  $I_n \sim n^{1/2}e^{bn}$  with  $b > 0$ . Therefore,  $\nu$  affects the value of  $\beta$  for which a group of size  $n$  can sustain an epidemic locally, but the behavior of  $I_n$  is unaltered compared to the  $\nu = 0$  case.

### 3.12.3 Mesoscopic localization

The form of Eqs. (3.34) and (3.35) is similar to Eqs. (3.17) and (3.13). It is then straightforward to obtain the asymptotic localization regimes as in Sec. 3.6.1 by investigating the behavior of  $\beta_c n_{\max}^{1-\nu}$ . Note that the scaling for the second term of the coupling is now

$$\frac{\langle n^{1-\nu}(n-1) \rangle}{\langle n \rangle} \sim \begin{cases} n_{\max}^{3-\gamma_n-\nu} & \text{if } \gamma_n + \nu < 3, \\ \ln n_{\max} & \text{if } \gamma_n + \nu = 3, \\ 1 & \text{if } \gamma_n + \nu > 3. \end{cases} \quad (3.37)$$

If  $\nu = 1$ , we always have  $\beta_c n_{\max}^{1-\nu} \rightarrow q < 1$  or 0, hence the epidemic is always *delocalized*. Therefore, let us focus on  $\nu \in [0, 1)$ . If  $\gamma_m \geq 3$ , we have  $\beta_c n_{\max}^{1-\nu} \rightarrow 1$  for all  $\gamma_n$ , as in the case  $\nu = 0$ , meaning that the outbreak is always *localized*. This is surprising, since  $\nu > 0$  increases the value of  $\beta$  for which a group of size  $n$  is able to sustain an epidemic locally. The reason is that  $\nu > 0$  also *decreases* the coupling between groups  $\Omega(p_n, g_m; \nu)$ , hence both effects cancel each other.

If  $2 < \gamma_m < 3$  and

- $2 < \gamma_n + \nu < 3$ , then

$$\beta_c n_{\max}^{1-\nu} \rightarrow \begin{cases} 0 & \text{if } 3 - \gamma_n + \alpha(3 - \gamma_m) > 1, \\ q < 1 & \text{if } 3 - \gamma_n + \alpha(3 - \gamma_m) = 1, \\ 1 & \text{if } 3 - \gamma_n + \alpha(3 - \gamma_m) < 1, \end{cases} \quad (3.38a)$$

- $\gamma_n + \nu = 3$ , then

$$\beta_c n_{\max}^{1-\nu} \rightarrow \begin{cases} 0 & \text{if } \alpha(3 - \gamma_m) + \nu \geq 1, \\ 1 & \text{if } \alpha(3 - \gamma_m) + \nu < 1, \end{cases} \quad (3.38b)$$

- $\gamma_n + \nu > 3$ , then

$$\beta_c n_{\max}^{1-\nu} \rightarrow \begin{cases} 0 & \text{if } \alpha(3 - \gamma_m) + \nu > 1, \\ q < 1 & \text{if } \alpha(3 - \gamma_m) + \nu = 1, \\ 1 & \text{if } \alpha(3 - \gamma_m) + \nu < 1. \end{cases} \quad (3.38c)$$

Again, we see that for  $2 < \gamma_n + \nu < 3$ , Eq. (3.38a), scaling the transmission rate with  $n^{-\nu}$  does not affect the localization regime in the  $(\gamma_m, \gamma_n)$  space. The effect becomes perceptible for  $\gamma_n + \nu \geq 3$ ,

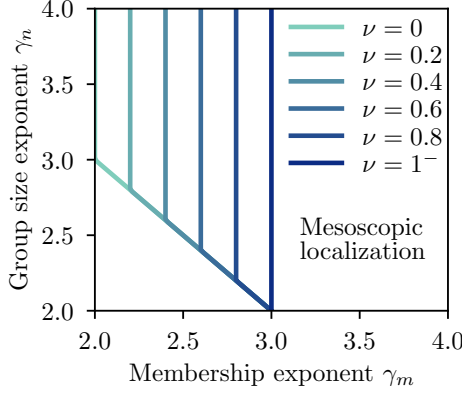


FIGURE 3.9 – Impact of scaling the transmission rate on the boundary separating the localized and delocalized regimes for power-law membership and group size distributions. We used  $m, n \in \{2, \dots, n_{\max}\}$ , hence  $\alpha = 1$ , and different values of  $\nu$  for the relation  $\beta \mapsto \beta n^{-\nu}$ . The boundaries are obtained from Eqs. (3.38).

when the coupling  $\Omega(p_n, g_m; \nu)$  is dominated by the first term depending solely on the membership distribution.

Figure 3.9 shows the impact of  $\nu > 0$  on the boundary separating the localized and delocalized regimes. The top portion of the boundary moves to higher values of  $\gamma_m$  as  $\nu$  is increased, reducing the size of the mesoscopic localization region. In the limit  $\nu \rightarrow 1^-$ , there still exists a nonvanishing portion of the parameter space allowing localization, i.e.,  $\gamma_m \geq 3$ . At  $\nu = 1$ , mesoscopic localization is impossible for all  $\gamma_m$  and  $\gamma_n$ , and thus there is no boundary.

### 3.13 Appendix : Localized portion of the bifurcation diagram

As  $\beta$  is increased beyond  $\beta_c$ , groups of smaller sizes can self-sustain the epidemic locally, until a point where the disease is present in all groups—the epidemic is not localized anymore. When the epidemic becomes delocalized, the global prevalence curve reaches an inflection point—the second derivative with respect to  $\beta$  turns negative because all groups sustain the epidemic and saturation effects become more important. This can be seen for instance in Fig. 3.10 around  $\beta/\beta_c = 4.3$ . But how do we define the range for  $\beta$  where the epidemic is considered localized, and how is this range affected by the structure?

An informal definition is to consider an epidemic localized for  $\beta \in [\beta_c, \beta^*]$ , where  $\beta^* \equiv \Omega(g_m, p_n; \nu)^{-1}$  acts as a *delocalization threshold*. Indeed, in the delocalized regime, we have that  $\beta_c \approx \Omega(g_m, p_n; \nu)^{-1}$ . This reinforces the interpretation of  $\Omega(g_m, p_n; \nu)$  as a structural *coupling* between the groups : for  $\beta \Omega(g_m, p_n; \nu) > 1$ , the disease is able to efficiently spread between groups, and the disease is sustained *collectively*.

It is analogous to the observation made in Ref. [195] that the epidemic threshold predicted by the heterogeneous mean-field theory seems to predict the delocalization threshold. In fact, if we take  $p_n = \delta_{n,2}$  (equivalent to having configuration model networks with degree distribution  $g_m$ ), the threshold



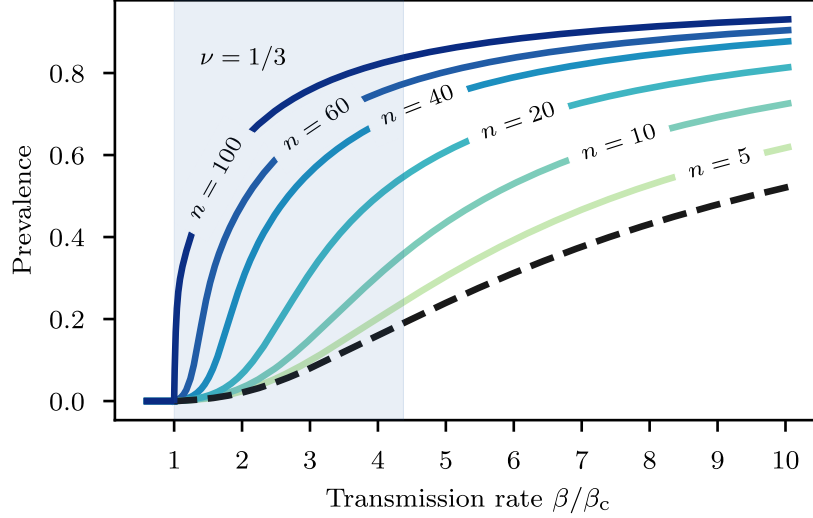


FIGURE 3.10 – Localized portion of the bifurcation diagram in the mesoscopic localization regime. The shaded region represents the localized portion, defined as  $[\beta_c, \Omega(g_m, p_n; \nu)^{-1}]$ . We used the same structure as in Fig. 3.3(e), but with  $\beta \mapsto \beta n^{-\nu}$ , where  $\nu = 1/3$ . The solid lines represent the group prevalence and the dashed line represents the global prevalence. Stationary state solutions were obtained from Eqs. (3.5a) and (3.6). The epidemic threshold  $\beta_c$  is the solution to Eq. (3.33).

$\beta^*$  is equivalent to the one predicted by pair heterogeneous mean-field theory [193], i.e.,

$$\beta^* = \frac{\langle m \rangle}{\langle m(m-1) \rangle}.$$

In Fig. 3.10, the shaded region highlights the localized portion of the bifurcation diagram. Note that the right-hand side of this region roughly corresponds to the inflection point of the global prevalence.

## Chapitre 4

# Localisation et mesures de contrôle

Article original :

### **Social Confinement and Mesoscopic Localization of Epidemics on Networks**

**Guillaume St-Onge**, Vincent Thibeault, Antoine Allard, Louis J. Dubé et Laurent Hébert-Dufresne

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

Référence : Phys. Rev. Lett. **126**, 098301 (2021) [267]

© 2021 American Physical Society (§4.3-4.7)<sup>1</sup>

---

1. Ces sections contiennent le contenu original de l'article. Celui-ci n'a été modifié que pour se conformer au format exigé par la Faculté des études supérieures et postdoctorales de l'Université Laval.

## 4.1 Avant-propos

L'article sur lequel est basé ce chapitre [267] fut publié de manière conjointe avec celui du chapitre précédent [266]. L'approche théorique et le contexte étant essentiellement les mêmes, une lecture préalable du chapitre 3 permettra au lecteur de lire plus rapidement les premières sections ci-dessous. Les symboles utilisés sont aussi les mêmes que ceux du tableau 3.1.

Ce chapitre pousse la réflexion plus loin et étudie l'incidence des interventions ciblées sur les groupes en relation avec le phénomène de localisation mésoscopique. Pour bien comprendre les motivations de cet article, il faut retourner au début de la pandémie de COVID-19. À ce moment, alors que les frontières se fermaient et que les autorités de santé publique dans chaque pays commençaient à discuter de confinement, il y eut plusieurs mesures mises en place pour restreindre les événements et les regroupements dépassant une certaine taille [187]. Mais que représente une taille acceptable pour un regroupement en temps de pandémie? Et comment modéliser l'effet de telles mesures? Dans ce chapitre, nous montrons que pour modéliser adéquatement l'incidence de ces interventions, il faut tenir compte de la localisation.

## 4.2 Résumé

Les mesures de contrôle pour endiguer les épidémies sont souvent centrées sur le comportement des individus. Il est beaucoup plus difficile d'avoir des recommandations précises en ce qui a trait aux événements et aux phénomènes collectifs en général, car les modèles standards ne prennent pas en compte la structure d'ordre supérieur permettant de décrire ces regroupements. En utilisant une description des processus de contagion sur réseaux d'ordre supérieur, nous étudions l'effet de l'annulation des événements dont la taille dépasse une valeur seuil. Nous observons que l'épidémie peut être contrôlée efficacement lorsque l'intervention est centrée sur les groupes plutôt que sur les individus directement. Ce résultat est lié à la localisation mésoscopique des processus de contagions, auquel cas la propagation s'effectue principalement à l'intérieur de certains groupes.

## 4.3 Abstract

Recommendations around epidemics tend to focus on individual behaviors, with much less efforts attempting to guide event cancellations and other collective behaviors since most models lack the higher-order structure necessary to describe large gatherings. Through a higher-order description of contagions on networks, we model the impact of a blanket cancellation of events larger than a critical size and find that epidemics can suddenly collapse when interventions operate over groups of individuals rather than at the level of individuals. We relate this phenomenon to the onset of mesoscopic localization, where contagions concentrate around dominant groups.

## 4.4 Introduction

Standard disease models reduce the complexity of epidemics to simple processes that provide useful insights. In fact, many of the key results of these models provide the foundation for our current understanding and forecasting of novel emerging epidemics [160–162]. The reduction of the complex to the simple is perhaps best embodied by the mass-action approximation [72]. This assumption essentially

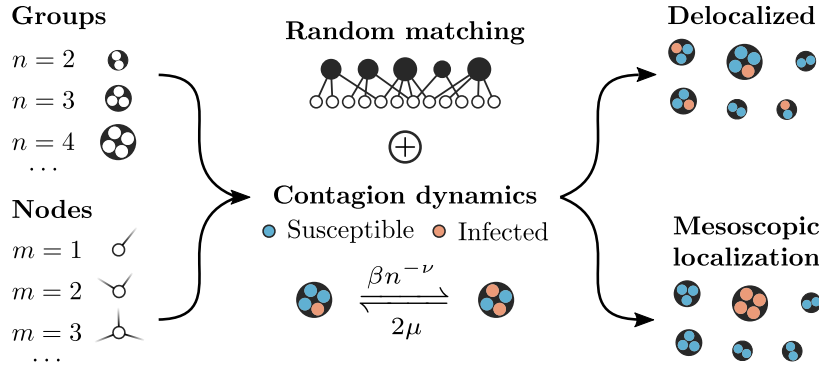


FIGURE 4.1 – Framework for contagions on higher-order networks. Nodes are assigned to  $m$  groups and groups are of various sizes  $n$ , distributed according to  $g_m$  and  $p_n$ . We consider a SIS dynamics where infected nodes transmit the disease in a group of size  $n$  at rate  $\beta n^{-\nu}$  with  $\nu \in [0, 1]$ , and recover at rate  $\mu$ . We characterize the phenomenon of mesoscopic localization, namely the concentration of infected nodes in large groups.

means that we are considering a randomly mixed population, ignoring household structures, social gatherings, and the different behaviors of different individuals. Mass-action models are thus seriously limited, since they focus on the average number of infections caused by each case, the basic reproduction number  $\mathcal{R}_0$  [12], and ignore the underlying heterogeneity [140]. There are also conceptual issues with the design of targeted interventions when relying on the mass-action assumption. Where should we target our interventions, and what should be their impact? In this chapter, we address these questions through higher-order contact patterns.

Network science provides a natural framework to go beyond the mass-action approximation by considering key features of the structure of contacts among individuals. The simplest generalization is perhaps the heterogeneous pair approximation—individuals are nodes categorized by their number of contacts and their state, and the contacts are distinguished by the states of the nodes involved [231]. At this level of sophistication, all pairwise contacts of a given state are still, *a priori*, equivalent.

One developing area in network science concerns dynamical processes on higher-order representation of networks, i.e., where the network is not simply a conglomerate of pairwise interactions but where interactions occur in a coordinated manner because of a higher-level organization (schools, households, events, etc.) [22]. For dynamics on higher-order networks, one straightforward generalization of the framework just described is the heterogeneous clique approximation [125]. Nodes are categorized by their state and *membership*, i.e., the number  $m$  of groups to which they belong. The groups are characterized by their size  $n$ , and the states of the nodes involved.

## 4.5 Group-based framework

Let us consider contagion processes on a simple version of higher-order networks, see Fig. 4.1. The network is characterized by  $g_m$ , the distribution for the memberships  $m$  of nodes and  $p_n$ , the distribution for the sizes  $n$  of groups. We use different heterogeneous distributions for both of them,  $g_m \propto m^{-\gamma_m}$  and  $p_n \propto n^{-\gamma_n}$ , with finite cutoffs  $m_{\max}$  and  $n_{\max}$  respectively.

For mathematical convenience, we use a susceptible-infected-susceptible (SIS) dynamics. However, our results have repercussions for a much broader class of dynamical processes, including susceptible-

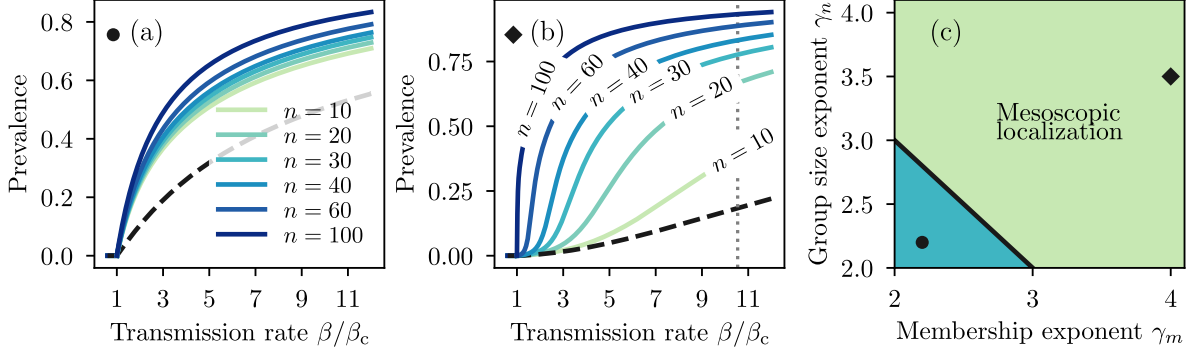


FIGURE 4.2 – Epidemic localization in networks with heterogeneous higher-order structures. We use  $\nu = 0$  for all panels and power-law distributions for the memberships  $g_m \propto m^{-\gamma_m}$  and group sizes  $p_n \propto n^{-\gamma_n}$ . (a)-(b) Solid lines represent the stationary group prevalence and dashed lines represent the stationary global prevalence. (a) For strongly coupled groups ( $\gamma_m = \gamma_n = 2.2$ ), we find a collective phase transition. (b) For weakly coupled groups ( $\gamma_m = 4, \gamma_n = 3.5$ ), we find a phenomenon of mesoscopic localization. While the global prevalence in the population can remain extremely low, larger groups can self-sustain the epidemic. The vertical dotted line is an estimation of the *delocalization* threshold, where the epidemic invades the whole network (see chapter 3). (c) Mesoscopic localization is the rule rather than the exception. The solid line,  $\gamma_m + \gamma_n = 5$ , separates the delocalized regime (blue area of strong group coupling) from the mesoscopic localization regime (green area of weak group coupling), obtained from Eqs. (4.8a) and (4.8b). The circle and diamond markers correspond to the networks in panels (a)-(b).

infected-recovered (SIR) dynamics where individuals develop immunity. Infected nodes transmit the disease to susceptible nodes belonging to a same group of size  $n$  at rate  $\beta n^{-\nu}$  with  $\nu \in [0, 1]$ , and recover at rate  $\mu$ . It is equivalent to a standard SIS model on networks formed of cliques [125, 218], but the edges have weights  $n^{-\nu}$ . We recover an unweighted network for  $\nu = 0$ . The parameter  $\nu$  tunes the strength of interactions within groups, which we assume to be decreasing with size. For instance, an individual in a workplace typically interacts with more people than at home, but interactions in a household are stronger.

We track  $s_m(t)$ , the probability for a node of membership  $m$  to be susceptible, and  $c_{n,i}(t)$ , the probability to observe  $i$  infected nodes within a group of size  $n$ . Their dynamics are described by the following coupled ordinary differential equations [125],

$$\frac{ds_m}{dt} = \mu(1 - s_m) - m r s_m, \quad (4.1a)$$

$$\begin{aligned} \frac{dc_{n,i}}{dt} = & \mu(i+1)c_{n,i+1} - \mu i c_{n,i} + (n-i+1)\{\beta n^{-\nu}(i-1) + \rho\}c_{n,i-1} \\ & - (n-i)\{\beta n^{-\nu}i + \rho\}c_{n,i}, \end{aligned} \quad (4.1b)$$

with the mean fields  $r(t)$  and  $\rho(t)$  defined as

$$r(t) = \frac{\sum_{n,i} \beta n^{-\nu} i (n-i) c_{n,i}(t) p_n}{\sum_{n,i} (n-i) c_{n,i}(t) p_n}, \quad (4.2a)$$

$$\rho(t) = r(t) \frac{\sum_m (m-1) m s_m(t) g_m}{\sum_m m s_m(t) g_m}. \quad (4.2b)$$

If we take a susceptible node and select a random group to which it belongs,  $r(t)$  is the mean infection rate associated to that group. Now if we pick a susceptible node in a group,  $\rho(t)$  is the mean infection

rate received from all external groups (i.e., excluding the one we picked the node from). Without loss of generality, we set  $\mu = 1$  hereafter.

An important feature of this framework is that Eq. (4.1b) is an approximate master equation : it describes the full range of possible states for groups of size  $n$ , while assuming a mean-field coupling between them. As we show in chapter 3, the agreement with Monte Carlo simulations is excellent. The global prevalence in the network—the average fraction of infected nodes—is then

$$I(t) = \sum_m [1 - s_m(t)] g_m , \quad (4.3)$$

and the group prevalence is

$$I_n(t) = \sum_i \frac{i}{n} c_{n,i}(t) . \quad (4.4)$$

## 4.6 Results

In Figs. 4.2(a) and (b), we show the stationary prevalence (global and within groups) for two different networks, obtained using Eq. (4.1). As expected from standard models, there exists an epidemic threshold  $\beta_c$  for the transmission rate below which epidemics cannot be sustained (see chapter 3 for an analytical expression of  $\beta_c$ ). Above  $\beta_c$ , the disease-free equilibrium of the dynamics becomes unstable, driving the epidemic to invade the network.

What is less expected are the local sequential transitions observed in the second panel [Fig. 4.2(b)]. For any value of the transmission rate, the outbreak thrives only in groups above a certain size. The epidemic is *self-sustained* locally, and the global prevalence reaches its highest growth rate with  $\beta$  well above the epidemic threshold, a defining feature of *smearred* phase transitions [118, 283]. This is reminiscent of certain infections, such as the bacteria *C. difficile*, mainly found in hospitals with large susceptible populations in close contact [197].

To get some insights on the emergence of this localization phenomenon, we examine the stationary group prevalence  $I_n^*$  near the absorbing state. Using a saddle-point approximation valid for large  $n$ , we obtain (chapter 3)

$$I_n^* \sim \begin{cases} \frac{1}{1 - \beta n^{1-\nu}} & \text{if } \beta < n^{\nu-1} \\ n^{1/2} (\beta n^{1-\nu})^n e^{-n+n^\nu/\beta} & \text{if } \beta \geq n^{\nu-1} . \end{cases} \quad (4.5)$$

For  $\beta > n^{\nu-1}$ , this implies  $I_n^* \sim n^{1/2} e^{bn}$  with  $b > 0$ . Therefore, if  $\beta_c \rightarrow n_{\max}^{\nu-1}$ , the group prevalence increases exponentially with  $n$  above the epidemic threshold, and the outbreak is *localized* in large groups, as observed in Fig. 4.2(b). In other words, the behavior of the epidemic threshold dictates whether or not localization is possible for a given network organization.

In chapter 3, we show that for power-law distributions of membership and group size, we have the following behavior

$$\beta_c^{-1} \sim \Omega(g_m, p_n; \nu) + n_{\max}^{1-\nu} , \quad (4.6)$$

where we define the *coupling* between groups as

$$\Omega(g_m, p_n; \nu) = \left( \frac{\langle m(m-1) \rangle}{\langle m \rangle} \right) \left( \frac{\langle n^{1-\nu}(n-1) \rangle}{\langle n \rangle} \right) . \quad (4.7)$$

Asymptotic analysis of Eqs. (4.6) and (4.7) in the limit  $n_{\max} \rightarrow \infty$  reveals the conditions for which  $\beta_c \rightarrow n_{\max}^{\nu-1}$ , i.e., the conditions necessary for a localized epidemic (chapter 3). They require  $\nu < 1$ , and are always met whenever  $\gamma_m \geq 3$ . If  $2 < \gamma_m < 3$ , the conditions are then satisfied only if

$$2 < \gamma_n + \nu < 3 \quad \text{and} \quad 3 - \gamma_n + \alpha(3 - \gamma_m) < 1, \quad (4.8a)$$

or

$$\gamma_n + \nu \geq 3 \quad \text{and} \quad \alpha(3 - \gamma_m) + \nu < 1, \quad (4.8b)$$

where  $\alpha \geq 0$  relates the two cutoffs  $m_{\max} \sim n_{\max}^\alpha$ . These conditions define *mesoscopic* localization and distinguish the *localized* regime from the *delocalized* regime. We give some examples with  $\nu = 0$  and  $\alpha = 1$  in Fig. 4.2(c).

More intuitively, the conditions (4.8a) and (4.8b) can be interpreted as the result of a competition between a collective and a local activation [91]. All pairs of  $(\gamma_m, \gamma_n)$  below the dark dividing line in Fig. 4.2(c) are associated with a strong group coupling ( $\Omega \gg n_{\max}^{1-\nu}$ ), whereas pairs above the line correspond to a weak group coupling ( $\Omega \ll n_{\max}^{1-\nu}$ ).

One important observation is that a large fraction of the structural parameter space  $(\gamma_m, \gamma_n)$  corresponds to the mesoscopic localization regime [green region in Fig. 4.2(c)], making it the rule rather than the exception. Moreover, the delocalized regime [blue region in Fig. 4.2(c)] is the parameter subspace where the underlying networks are dense, i.e., where the average number of contacts of a node, proportional to  $\langle n(n-1) \rangle$ , diverges in the asymptotic limit. Since real-world networks are generally sparse [69, 220], it is reasonable to assume mesoscopic localization may occur in many real-world networks with a higher-level organization. The results of Fig. 4.2 extend nicely to cases with  $\nu > 0$  (see chapter 3, Sec. 3.12) where the localized regime still dominates in a large portion the structural parameter space.

It is worth noting that we only observe localization for a certain portion  $\beta \in [\beta_c, \beta^*]$  of the bifurcation diagram, where  $\beta^*$  is the *delocalization* threshold [dotted line in Fig. 4.2(b)]. In chapter 3, we show how to estimate  $\beta^*$  and discuss other details of the localization regimes, notably the effects of finite cutoffs.

Perhaps most surprising about mesoscopic localization is how strong the effect can be. Even at low overall prevalence, we observe intense but local outbreaks in large groups. This simple observation justifies targeted interventions on these groups, analogous to school closures and the cancellation of large social or professional events. While such closures may seem excessive given the low prevalence found in the general population, these compact organized groups are where most infections will occur.

We now focus on targeted interventions on large groups, modeled by simply forcing a hard cutoff  $n_{\max}$  on the distribution  $p_n$ . By using a cutoff instead of immunizing the groups, we preserve the membership  $m$  of nodes; they will simply belong to groups of smaller sizes.

To compare the effectiveness of interventions across different networks and parametrizations, we define an intervention strength (IS) as the fraction of the total edge weights that have been removed. If  $p_n$  and  $\tilde{p}_n$  are, respectively, the group size distribution before and after the intervention, then

$$\text{IS} = 1 - \frac{\sum_n n^{1-\nu}(n-1)\tilde{p}_n}{\sum_n n^{1-\nu}(n-1)p_n}. \quad (4.9)$$

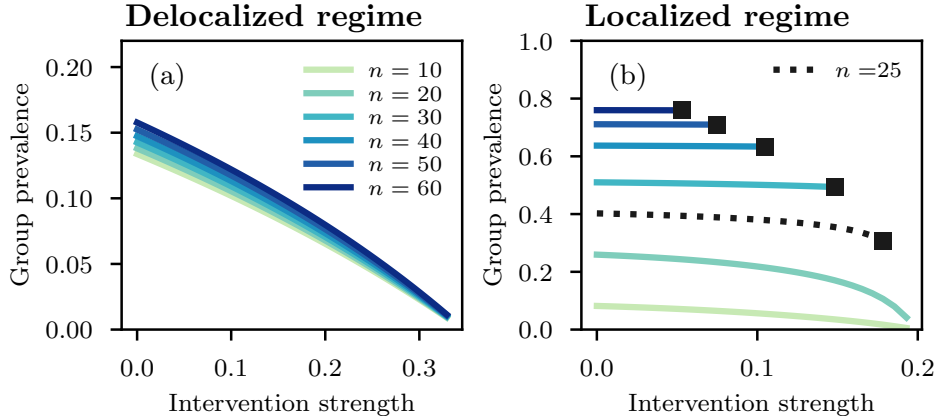


FIGURE 4.3 – Local impact of structural interventions in delocalized and localized epidemics. (a)-(b) Prevalence within groups of size  $n$  against the intervention strength [Eq. (4.9)]. We use the networks from Figs 4.2(a) and 4.2(b) with transmission rates adjusted to have similar global prevalence for both regimes without intervention. (a) In the delocalized regime, using  $\beta \approx 0.0041$ , we find a similar benefit of the intervention among all groups. (b) In the mesoscopic localization regime, using  $\beta = 0.07$ , we find a different story. Large groups that have not been removed by the intervention are barely affected, until the intervention is strong enough to cause a global collapse of the epidemic. Square markers indicate when groups of a particular size are removed.

In Fig. 4.3, we show the local impact on the group prevalence for such structural interventions. For networks in the delocalized regime [Fig. 4.3(a)], the intervention appears to reduce the risk of infections at the individual (node) level. As we decrease  $n_{\max}$  (and therefore increase the intervention strength), the local prevalence within all groups decreases gradually and homogeneously until an epidemic threshold is reached. This is similar to traditional models where interventions reduce  $\mathcal{R}_0$  in a distributed, mass-action way.

In the localized regime [Fig. 4.3(b)], the intervention has a very different impact. Individuals that would have interacted in groups of size greater than  $n_{\max}$  are spared by the intervention, but the large groups of sizes below  $n_{\max}$  appear unaffected by the intervention. The main point here is that a local outbreak in certain organized gatherings (e.g., mass transit in urban centers, cruise ships) can persist despite interventions elsewhere. However, once the intervention is strong enough [roughly when  $n_{\max}$  is below 25 in Fig. 4.3(b)], further interventions cause a rapid collapse of the epidemic.

Let us now look at the global impact of these interventions in Fig. 4.4. It is important to note that while we have assumed a SIS dynamics, our results hold for a SIR dynamics as well [see the similarity of Figs. 4.4(a) and 4.4(b)], which is generally a more realistic model for epidemics. The reason is that smeared phase transitions occur for this type of process as well [118].

In Fig. 4.4(a), we observe that the global impact of interventions in the delocalized regime is again similar to mass-action models—the prevalence decreases approximately linearly with the intervention strength. In contrast, in the localized regime, the intervention has a rapid nonlinear effect in reducing the global prevalence as its strength is ramped up.

In both Figs. 4.3(b) and 4.4(a), it is surprising that the removal of the largest groups does not produce the largest decrease in prevalence, since these are the ones in which we expect most nodes to be infected. Even less clear is what drives the sudden collapse and if we should expect this behavior for



all localized epidemics.

To clarify the situation, we need to explore the regimes between the delocalized and the *strongly* localized regimes displayed in Figs. 4.3(b), 4.4(a) and 4.4(b). We break down these regimes in three parts in Fig. 4.4(c) to simplify the discussion : (1) strongly localized, (2) weakly localized, and (3) delocalized. In the strongly localized regime (1), large groups have a high prevalence, and act as independent entities that are barely affected by interventions elsewhere [as in Fig. 4.3(b)]. In the weakly localized regime (2), the disease still thrives in large groups, but they are not isolated from one another—interventions in one group now affect the others as well. This is the regime where interventions are most effective. Finally, in the delocalized regime (3), groups act as a whole, but the infection does not thrive in any of them, leading to a lower effectiveness of targeted interventions.

The easiest way to interpolate between the two extremes is to tune the group interaction strength through  $\nu$ . As shown in Fig. 4.4(d), the initial impact of interventions changes as  $\nu$  is increased. With  $\nu = 0.6$  for instance, the epidemic is weakly localized, and now the removal of the largest groups produce the largest decrease in prevalence. Further interventions become eventually less effective, as the intervention itself causes the epidemic to shift to a delocalized regime.

More generally, the nonlinear decrease of the global prevalence as a function of the intervention strength is explained by transitions between different localization regimes, as illustrated in Fig. 4.4(e). The sudden collapse for interventions on strongly localized epidemics is thus the result of a shift to a weakly localized regime, where interventions become much more effective.

## 4.7 Conclusion and perspectives

Altogether, the lesson from Figs. 4.3 and 4.4 is that, just as we take heterogeneity of individual risks into account when preferentially vaccinating individuals, we should take heterogeneity of group risks into account when designing interventions. While pathogens operate at the scale of individuals, epidemics themselves interact with our entire social network, which has a modular, hierarchical, higher-order structure. Since we expect real epidemics to experience localization effects, we should aim to leverage their sudden collapse when designing structural interventions, and account for this emergent feature in our models.

Over the last few years, dynamics on higher-order representation of networks have shown time and time again that intuition built from simpler models does not always hold in more complex scenarios, with examples ranging from competitive dynamics [113] to social contagion [142]. Localization of dynamics over higher-order structures is yet another addition to this list. Our work thus paves the way for a more comprehensive analysis of structural interventions on networks.

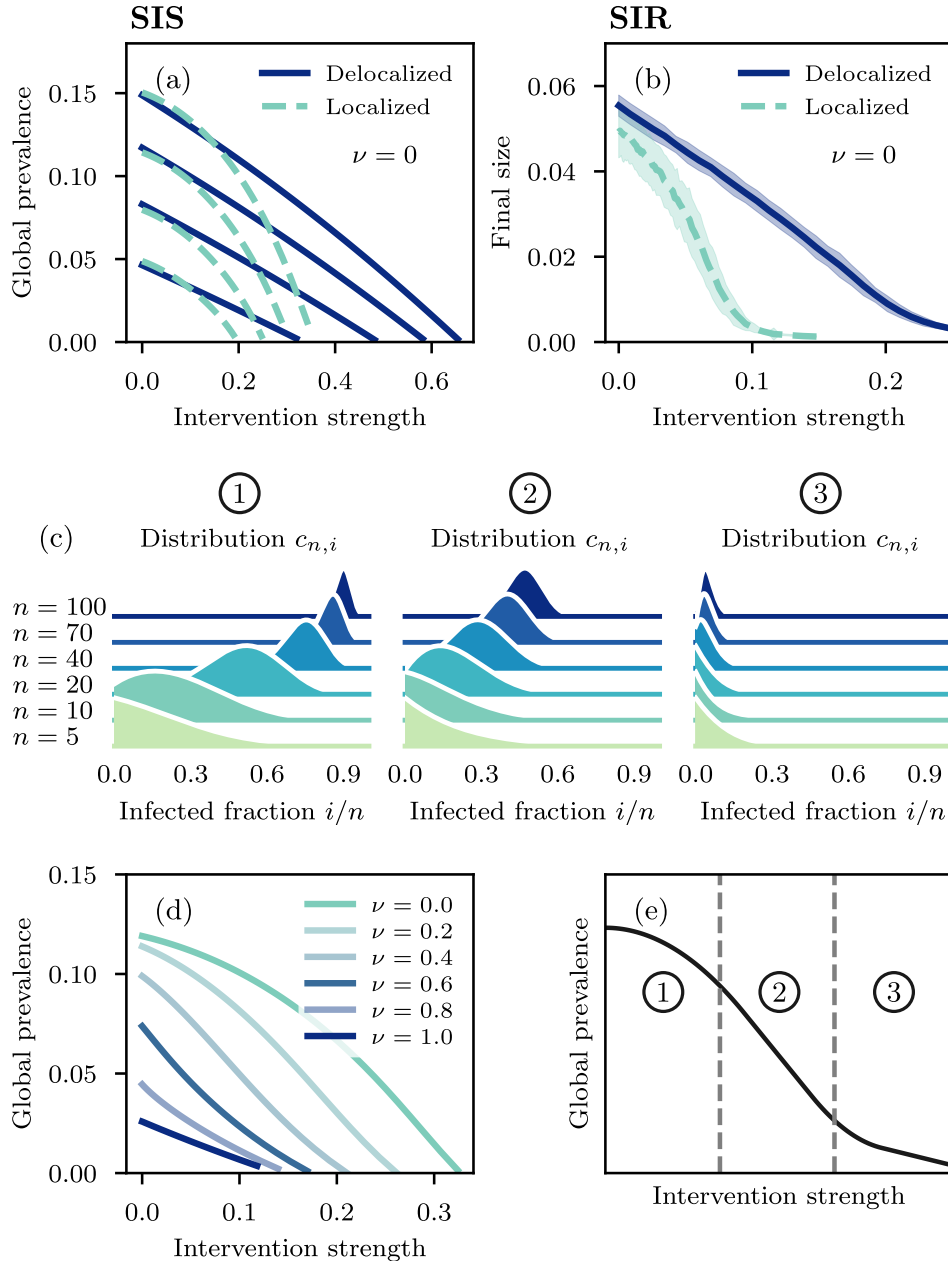


FIGURE 4.4 – Global impact of structural interventions in different localization regimes of SIS and SIR dynamics. (a) We compare the global prevalence of epidemics against the intervention strength [Eq. (4.9)] using the same networks as in Fig. 4.3. We use  $\beta \in \{0.004, 0.005, 0.006, 0.007\}$  and  $\beta \in \{0.07, 0.09, 0.11, 0.13\}$  for the delocalized and localized regimes. (b) A similar behavior is observed for the final size (fraction of recovered nodes at  $t \rightarrow \infty$ ) in simulations of the SIR dynamics. We use the algorithm developed in Appendix A. We generate 1000 networks of size  $N = 10^6$  and run  $10^4$  SIR simulations on each network, starting with a single random infected node, with  $\beta = 0.0036$  in the delocalized regime and  $\beta = 0.083$  in the localized regime. We only kept macroscopic outbreaks, with a final size over  $10^{-3}$ . Solid and dashed lines are sample means and the shaded regions represent twice the standard deviation. (c) Three examples of  $c_{n,i}$  distributions for (1) strongly localized, (2) weakly localized, and (3) delocalized epidemics, using  $\nu$  equal 0, 0.6 and 1 respectively. We used networks with  $\gamma_m = \gamma_n = 3.5$  and  $\beta = 0.3/\langle k \rangle$ , where  $\langle k \rangle$  is the average weighted degree before interventions. (d) We look at the impact of interventions for intermediate localization regimes by varying  $\nu$  with the same networks as in (c). (e) Schematic representation of how the impact of interventions varies as it passes through different localization regimes.

Troisième partie

Contagions complexes

## Chapitre 5

# Universalité des contagions non linéaires en loi de puissance

Article original :

### **Universal Nonlinear Infection Kernel from Heterogeneous Exposure on Higher-Order Networks**

**Guillaume St-Onge**, Hanlin Sun, Antoine Allard, Laurent Hébert-Dufresne et Ginestra Bianconi

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

Référence : Phys. Rev. Lett. **127**, 158301 (2021) [265]

© 2021 American Physical Society (§5.3-5.12)<sup>1</sup>

---

1. Ces sections contiennent le contenu original de l'article. Celui-ci n'a été modifié que pour se conformer au format exigé par la Faculté des études supérieures et postdoctorales de l'Université Laval et pour avoir une notation et une terminologie cohérentes avec les autres chapitres.

## 5.1 Avant-propos

Dans les deux chapitres précédents, nous nous sommes concentrés sur les contagions simples, dont le taux d'infection d'un noeud est proportionnel au nombre de noeuds infectés dans son voisinage. C'est l'hypothèse sous-jacente à la grande majorité des modèles pour la transmission des maladies infectieuses [12]. De manière implicite, cela suppose que la transmission est fondamentalement le résultat d'une interaction dyadique : un individu contagieux infecte un individu susceptible avec une certaine probabilité.

Que ce soit la transmission d'un virus ou d'une bactérie, une *dose infectieuse minimale*<sup>2</sup> — un nombre de bactéries ou de particules virales — est requise (en moyenne) pour que la maladie se développe chez un individu [15, 171, 292]. Un individu accumule une dose infectieuse en étant en *contact* avec des individus infectés ; la nature de ce qui constitue un contact varie bien entendu avec le mode de transmission de l'agent pathogène. Par conséquent, si un individu est exposé à plusieurs individus contagieux, et que la dose reçue par chacun est insuffisante, mais que la dose combinée l'est, considérer les interactions d'ordre supérieur devient important pour bien modéliser la propagation. C'est l'hypothèse que nous explorons dans ce chapitre.

En particulier, nous combinons le concept de dose infectieuse minimale, la représentation en réseau d'ordre supérieur et une *durée de contact hétérogène* (voir la section 1.4.2). Afin de combiner tous ces éléments, nous modélisons le processus stochastique en temps discrets, avec  $t \in \mathbb{N}_0$ . L'interprétation et la description mathématique sont toutefois très analogues à ce qui a été présenté aux chapitres précédents. Par exemple, la probabilité qu'un noeud se fasse infecter à chaque temps discret peut être interprétée comme un taux d'infection.

L'article associé à ce chapitre est le fruit d'une collaboration scientifique avec un autre groupe de recherche, donc la perspective est légèrement différente de celle des autres chapitres. Les groupes (hyperliens) sont associés à des *environnements* fréquentés par un certain nombre d'individus. Aussi, plutôt que de référer au nombre de noeuds infectés dans un groupe,  $i$ , l'attention est portée sur la fraction des *voisins* infectés,  $\rho = i/(n - 1)$ , pour un noeud dans un groupe de taille  $n$ . Pour mieux harmoniser ce chapitre avec le reste de la thèse, la notation et la terminologie utilisées ici sont différentes de celles de l'article original.

Symbole	Description
$n$	Taille de groupe : nombre de noeuds dans un groupe (environnement)
$m$	Appartenance : nombre de groupes auxquels un noeud appartient
$\tau$	Durée de la participation dans un groupe (environnement)
$p_n$	Distribution des tailles de groupe
$g_m$	Distribution des appartenances
$\wp(\tau)$	Densité de probabilité pour la durée de participation

2. En français, le terme approprié semble être « dose infectieuse », mais le terme en anglais est « minimal infective dose ». Pour éviter toute confusion avec la partie anglaise, nous utilisons le terme « dose infectieuse minimale », et « dose » ou « dose infectieuse » pour le nombre de particules virales ou de bactéries reçus par un individu.

$\alpha$	Exposant de la loi de puissance $\wp(\tau) \propto \tau^{-\alpha-1}$
$i$	Nombre de noeuds infectés dans un groupe
$\varrho$	Fraction des voisins infectés dans un groupe, soit $i/(n-1)$ pour un groupe de taille $n$
$\kappa$	Dose infectieuse (particules virales ou bactéries) reçue lors de la participation à un environnement
$K$	Dose infectieuse minimale pour qu'un individu développe la maladie
$\lambda$	Dose infectieuse moyenne reçue dans un environnement
$\beta$	Taux d'accumulation de la dose infectieuse dans un environnement, donc $\lambda \propto \beta$
$\pi(\kappa; \lambda)$	Densité de probabilité associée à la dose infectieuse $\kappa$ reçue dans un environnement
$\mathcal{T}$	Temps caractéristique pour que le système immunitaire élimine toute dose infectieuse reçue
$\tau_c$	Temps caractéristique qu'un individu doit passer dans un environnement pour devenir infecté
$\theta_n(\varrho)$	Probabilité qu'un noeud devienne infecté suite à la participation à un environnement de taille $n$ avec une fraction $\varrho$ des autres noeuds qui sont infectés
$\nu$	Exposant résultant de l'analyse asymptotique $\theta_n(\varrho) \sim \tau_c^{-\nu} \propto \varrho^\nu$
$s_m(t)$	Fraction des noeuds d'appartenance $m$ qui sont susceptibles au temps $t$
$I(t)$	Prévalence : fraction des noeuds qui sont infectés au temps $t$

TABLEAU 5.1 – Glossaire des symboles importants utilisés au chapitre 5.

## 5.2 Résumé

La présence simultanée des individus dans différents lieux est une composante primordiale de la transmission des maladies infectieuses à l'intérieur d'un réseau social. Pourtant, les modèles épidémiologiques standards ne prennent pas en considération l'étendue de la complexité liée à l'exposition à plusieurs individus infectés. (1) Ils négligent la structure d'ordre supérieur des contacts entre individus, lesquels se produisent dans différents endroits, comme au travail, dans un restaurant, à l'intérieur des ménages, etc. (2) Ils supposent que le taux d'infection varie linéairement avec le nombre de contacts infectés auxquels un individu est exposé. Pour pallier ces limitations, notre modèle utilise un hypergraphe pour représenter la structure sociale, ce qui permet de prendre en compte l'hétérogénéité des lieux et l'hétérogénéité liée aux individus fréquentant ces lieux. En combinant un temps d'exposition hétérogène et le concept de dose infectieuse minimale pour qu'un individu devienne infecté, nous démontrons qu'il existe une relation non linéaire universelle entre le taux d'infection et le nombre de contacts potentiels avec des individus infectés. Cette relation non linéaire remet en question les propriétés typiques des modèles épidémiologiques : la transition de phase du système peut être discontinue avec une boucle d'hystérésis, et une épidémie peut croître super-exponentiellement.

### 5.3 Abstract

The collocation of individuals in different environments is an important prerequisite for exposure to infectious diseases on a social network. Standard epidemic models fail to capture the potential complexity of this scenario by (1) neglecting the higher-order structure of contacts that typically occur through environments like workplaces, restaurants, and households, and (2) assuming a linear relationship between the exposure to infected contacts and the risk of infection. Here, we leverage a hypergraph model to embrace the heterogeneity of environments and the heterogeneity of individual participation in these environments. We find that combining heterogeneous exposure with the concept of minimal infective dose induces a universal nonlinear relationship between infected contacts and infection risk. Under nonlinear infection kernels, conventional epidemic wisdom breaks down with the emergence of discontinuous transitions, superexponential spread, and hysteresis.

### 5.4 Introduction

Mathematical models of epidemics play an increasingly important role in public health efforts and pandemic preparedness [242]. By providing insights on the interplay of the biological and sociological aspects of epidemics, models can test potential interventions *in silico* and suggest potential outcomes [80]. However, large-scale forecasting comparisons show that statistical models often outperform mechanistic models that make assumptions about spreading dynamics [33].

In this chapter, we look at the interplay of two commonly used assumptions in disease models : Random mixing and the linearity between infection risk and exposures to infected individuals. In almost all disease models, doubling the number of contacts between susceptible and infectious individuals doubles the risk of infection for the susceptible individuals. Some past work in mathematical biology has considered nonlinear infection rates [129, 183], but these models are rarely used in practice. Other fields such as sociology do often consider generalized contagion models, often dubbed complex contagions [51, 176]. In these complex contagions, having a nonlinear relationship between infection rate and sources of infection allows the model to consider mechanisms such as social reinforcement [50], where a set of multiple exposures can have more impact than the mere sum of unique exposures.

The mathematical convenience of assuming random mixing when modeling infectious diseases comes at the price that all contacts between susceptible and infectious individuals are effectively equivalent. This assumption has often been lifted using heterogeneous mathematical models where individuals are distinguished by some individual features such as their intrinsic susceptibility or reaction to the infection [41, 84], relaxing the mass-action assumption directly [127, 128], or by specifying an underlying contact network [231, 233].

Moreover, the linearity assumption says that all increments in the total exposure to infectious individuals (measured for example as a viral inoculum) are equivalent. Evidence associated with the minimal infective dose of different infectious diseases shows that not all exposures are equal, and that some minimal dose might be required for an infection to likely occur. More precisely, the  $ID_{50}$  value is a measure of the dose needed to cause an infection in 50% of individuals. These concepts are needed because our immune system is usually able to handle microscopic challenges from viruses and bacteria alike. While an infective dose of tuberculosis might only require between 1 and 5 bacteria [15], some enterics might require up to  $10^9$  pathogenic particles [171], and others like common respiratory infec-

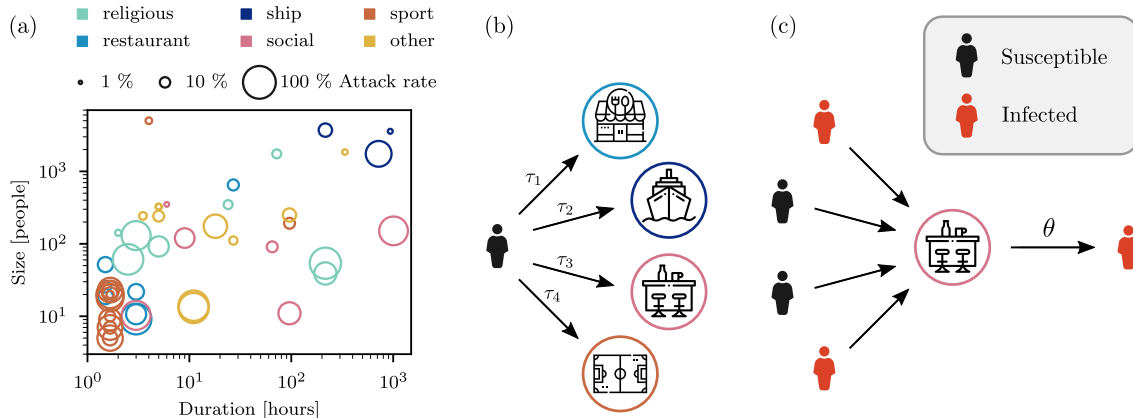


FIGURE 5.1 – Modeling contagions and superspreading events through higher-order networks. (a) Scatter plot of superspreading events of COVID-19 where the number of people involved (size), the duration of the event, and the resulting proportion of infected individuals (attack rate) are all available (extracted from Refs. [175, 272]; see Sec. 5.12). (b),(c) Framework for contagions on hypergraphs<sup>3</sup>, where the size  $n$  of the groups (environments), the memberships  $m$  of the nodes (individuals), and the participation time to the environment  $\tau$  are all heterogeneous, distributed according to  $p_n$ ,  $g_m$ , and  $\varphi(\tau)$ , respectively. (b) At each time step  $t$ , an individual participates for a time  $\tau$  (drawn independently) to each environment. (c) An individual gets infected with probability  $\theta_n(\varrho)$  in the environment at time step  $t$ , which depends on the size  $n$  and the fraction infected  $\varrho$ .

tions still require further study [292]. There are indeed multiple different physical mechanisms behind immune evasion—for example some airborne viruses need to find their receptors on lung epithelial cells, while some bacteria might instead require interaction with the immune system [97]. These mechanisms are reviewed in Refs. [44, 92, 97, 136, 179], and all of them combine to determine the ID50 of specific pathogens. Likewise, the decay or clearing rates of pathogens in noninfectious courses can also vary a lot, potentially requiring days for bacteria to hours for airborne viral infections. For example, mathematical models for the pathogenesis of SARS-CoV-2 or influenza A use decay rates of the order of 7–18 hours but empirical estimates vary wildly (see Ref. [77] and [23] and references therein).

To study the effect of simultaneously relaxing these two assumptions, we consider a social structure where individuals attend a certain number of environments such as work places, gyms, or supermarkets. This division of contact structure in environments is motivated by the known role of superspreading events, which are, for example, critical to the ongoing spread of COVID-19 [8, 29, 79, 173, 184, 200, 222, 295]. While variations at the individual level are often used to explain superspreading [185], we focus here on the variability of environments and of temporal patterns [48, 49, 134, 152, 153, 270, 304] at the *group* level, which undoubtedly affects epidemics [8], especially when a certain exposure within a certain time window is needed to confidently spark an infection. Interestingly, available case data highlight how there is no expected size or duration for such events. Transmission is highly context dependent on the settings (e.g., ventilation) and activity (e.g., singing or shouting) such that the resulting superspreading events are heterogeneous in size, duration and attack rate, as shown in Fig. 5.1(a). Higher-order contact structures and heterogeneous temporal patterns are therefore key ingredients for more realistic models of spreading dynamics.

3. Icons made by Freepik from www.flaticon.com.



Mathematically, we represent the contact structure as a hypergraph [22, 27, 277] where each environment is described by a group connecting  $n$  nodes (individuals) and where each node is incident to  $m$  groups. All groups of a same size  $n$  are considered equivalent, although this assumption is relaxed in Appendix 5.8 to consider additional sources of heterogeneity. To model heterogeneous temporal patterns, we consider a discrete-time process, where at each time step  $t = 1, 2, \dots$ , we draw for each individual a *participation time*  $\tau \in [1, \tau_{\max}]$  for each environment to which they are connected [Fig 5.1(b)]. The time steps correspond to fixed temporal windows of size  $\tau_{\max}$ , during which susceptible individuals can get infected through their participation to environments.

We first study the impact of the spatiotemporal collocation patterns on the *infection kernel*  $\theta_n(\varrho)$ , the probability of getting infected in an environment of size  $n$  when a fraction  $\varrho$  of the other participants are infectious [Fig. 5.1(c)]. We then analyze the properties of the resulting contagion process.

## 5.5 Universal infection kernel from heterogeneous exposure

Let us consider a susceptible individual participating to an environment of size  $n$  for a duration  $\tau$ , where a fraction  $\varrho$  of the other participants are infectious. During this exposure period, some of the other  $n - 1$  individuals might participate to the environment as well. We assume that the considered individual receives an infective dose  $\kappa \in [0, \infty)$  from the infectious individuals, distributed according to  $\pi(\kappa; \lambda)$ , where  $\lambda \equiv \langle \kappa \rangle$ . A reasonable assumption is that the mean dose received is proportional to the time spent in the environment and to the proportion of infectious people,  $\lambda = \beta f(n) \tau \varrho$ , where  $\beta$  is a rate of dose accumulation and  $f(n)$ , unitless, modulates the typical number of contacts in environments frequented by  $n$  individuals.

While this is not a strict requirement for our results to hold (see Appendix 5.8), we further assume that the random variable for the dose can be written as  $\kappa = \lambda u$ , where  $u$  is a random variable that is independent of  $\lambda$ . In this case,  $u$  can be seen as an intrinsic property of the contagion process—determined by rates of viral shedding, diffusion in the environment, variability of human interactions, etc.—while  $\lambda$  acts as a *scale* parameter, i.e.,  $\pi(\kappa; \lambda) = \pi(\kappa/\lambda; 1)/\lambda \equiv \pi(\kappa/\lambda)/\lambda$ .

To incorporate the concept of minimal infective dose, we assume that an individual develops the disease if  $\kappa > K$ , a perspective analogous to standard threshold models [75, 112, 288] and related to the assumption that successful host invasion necessitates multiple attempts by the pathogen [13]. The probability of getting infected in the environment is then

$$\bar{\Pi}(K/\lambda) = \int_{K/\lambda}^{\infty} \pi(\kappa) d\kappa. \quad (5.1)$$

The infection kernel  $\theta_n(\varrho)$  is calculated by averaging  $\bar{\Pi}(K/\lambda)$  over  $\varphi(\tau)$ . We focus here on the case of heterogeneous exposure periods modeled with a Pareto distribution  $\varphi(\tau) = C_\alpha \tau^{-\alpha-1}$ , where  $C_\alpha$  is a normalization constant,  $\alpha > 0$ , and  $\tau \in [1, \tau_{\max}]$ . However, for our dose mechanism to be well defined, we can only average over participation times  $\tau \in [1, \mathcal{T}]$ , where  $\mathcal{T} \leq \tau_{\max}$  is the “clearing window”, i.e., the characteristic time for the immune system to get rid of any dose  $\kappa < K$ . If we assume that this clearing window is sufficiently large compared to  $\tau_c \equiv K/\beta f(n) \varrho$ , the characteristic time to get infected in the environment, we can neglect events where  $\tau \geq \mathcal{T}$  as they do not contribute significantly to the infection kernel (see Appendix 5.8). We therefore redefine our support as  $\tau \in [1, \mathcal{T}]$  and the

infection kernel is

$$\begin{aligned}\theta_n(\varrho) &= \int_1^{\mathcal{T}} \bar{\Pi}(\tau_c/\tau) \wp(\tau) d\tau \\ &= \frac{C_\alpha}{\alpha} \left[ \bar{\Pi}(\tau_c) - \bar{\Pi}(\tau_c/\mathcal{T}) \mathcal{T}^{-\alpha} + \tau_c^{-\alpha} \int_{\tau_c/\mathcal{T}}^{\tau_c} \pi(y) y^\alpha dy \right].\end{aligned}\tag{5.2}$$

When  $1 \ll \tau_c \ll \mathcal{T}$  and  $\pi(y)$  decreases faster than  $y^{-\alpha-1}$ , then the integral on the right converges to a constant, the term in  $\mathcal{T}^{-\alpha}$  can be neglected, and  $\bar{\Pi}(\tau_c) \ll \tau_c^{-\alpha}$ , which implies

$$\theta_n(\varrho) \simeq D_\alpha \tau_c^{-\alpha} \propto \varrho^\alpha,\tag{5.3}$$

where  $D_\alpha$  is some constant. This form of infection kernel is *universal*, driven by temporal patterns, and does not depend on the value of  $K$  (given  $K > 0$ ) or on the particular form of  $\pi$ . We illustrate it in Fig. 5.2(a) using an exponential for the dose distribution—other cases such as the Weibull and the Fréchet distribution are presented in Appendix 5.8.

Let us stress that the condition  $1 \ll \tau_c \ll \mathcal{T}$  is not restrictive. On the timescale on which we measure the exposure periods,  $\tau_c \gg 1$  implies that the contagion does not spread too fast, because otherwise the whole population would be rapidly infected, while  $\mathcal{T} \gg \tau_c$  suggests that it is transmissible before the immune system is able to clear the dose received.

More broadly, we do not need to assume that  $\pi(y)$  decreases faster than  $y^{-\alpha-1}$ , nor that  $\lambda$  is a scale parameter for  $\kappa$ . In fact,  $\pi(\kappa; \lambda)$  does not even need to be a continuous distribution and  $\wp(\tau)$  could be asymptotically power law for large  $\tau$ . In this more general context, we still recover a universal infection kernel  $\theta_n(\varrho) \sim \varrho^\nu$  in most cases, but  $\nu$  is not always directly equal to  $\alpha$  (See Appendix 5.8). Linear infection kernels ( $\nu = 1$ ) are recovered only in some specific cases—for instance when  $\alpha = 1$  or when we use a Poisson distribution for  $\kappa$  and  $K = 1$ . From now on, we make abstractions of the underlying distributions  $\pi(\kappa; \lambda)$  and  $\wp(\tau)$  and focus on the resulting effective model parameterized by  $\nu$ .

## 5.6 Epidemic spreading with nonlinear infection kernel

We now illustrate the consequences of nonlinear infection kernels for contagion on hypergraphs. To simplify the mathematical analysis, we consider a Susceptible-Infectious-Susceptible model. At each time step, an infected node becomes susceptible with probability  $\mu$ , while a susceptible node gets infected through a group of size  $n$  with probability  $\theta_n$ . If a node is part of multiple groups, we assume that the probability of infection through each group is independent. This is reasonable if  $m_{\max} \mathcal{T} \leq \tau_{\max}$  for a maximal membership  $m_{\max}$ . For instance, if  $\mathcal{T}$  is of the order of a few hours,  $\tau_{\max}$  a week, and an individual participates to an environment once a day, the night allows the immune system to clear any dose  $\kappa < K$  accumulated the day before.

To obtain some analytical insights, we introduce a degree-based mean-field theory [231], an approximation equivalent to consider an infinite size *annealed* hypergraph, where at each time step, the connections between nodes and groups are shuffled. We validate our approach with Monte Carlo simulations on *quenched* hypergraphs in Appendix 5.11.

With the mean-field approximation, the marginal probability for a node to be susceptible at time  $t$  only depends on its membership, which we note  $s_m(t)$ . The global prevalence is then simply  $I(t) =$

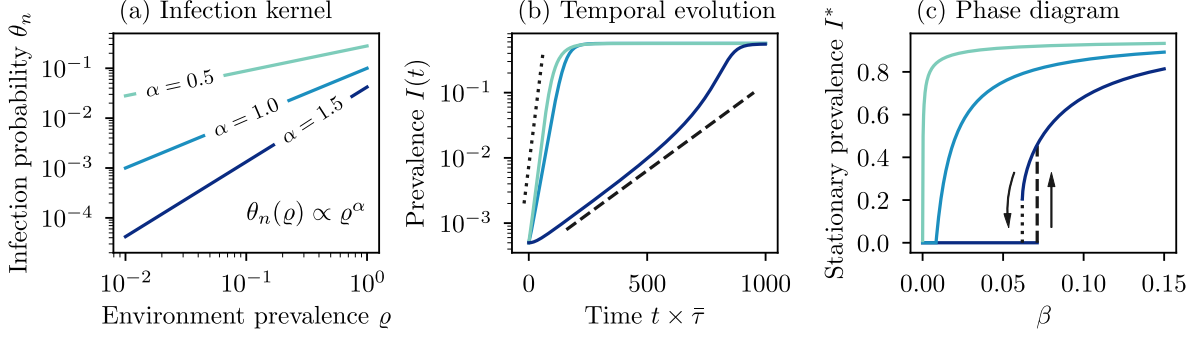


FIGURE 5.2 – Properties of contagions with nonlinear infection kernels induced by heterogenous exposure. We use an exponential dose distribution  $\pi(\kappa; \lambda) \propto e^{-\kappa/\lambda}$  with a power-law distribution of participation time  $\varrho(\tau) \propto \tau^{-\alpha-1}$ , a clearing window  $\mathcal{T} \rightarrow \infty$ , and  $f(n) = 1$ . (a) Effective infection kernel using  $\beta = 0.1$ . The infection probability has a power law scaling  $\theta_n(\varrho) \sim \varrho^\alpha$ . (b)-(c) We use Poisson distributions for both  $g_m$  and  $p_n$ , with  $\langle m \rangle = 5$  and  $\langle n \rangle = 10$ , and set  $\mu = 0.05$ . We use Eqs. (5.4)–(5.6) to evolve the system. (b) Supralinear kernels  $\nu > 1$  lead to a superexponential growth for the global prevalence  $I(t)$ . We use  $\beta = 5 \times 10^{-4}$ ,  $\beta = 0.025$  and  $\beta = 0.077$  for  $\nu = 0.5$ ,  $\nu = 1$  and  $\nu = 1.5$  respectively.  $\bar{\tau}$  is the median exposure period. (c) The phase diagram for stable solutions in the stationary state ( $t \rightarrow \infty$ ) can be continuous or discontinuous with a bistable regime. Sublinear and linear kernels  $\nu \leq 1$  lead to a continuous phase transition, and the invasion threshold  $\beta_c$  vanishes for  $\nu \rightarrow 0$ . Supralinear kernels  $\nu > 1$  can lead to a discontinuous phase transition with a bistable regime.

$\sum_m [1 - s_m(t)] g_m$  and the evolution of the system is described by [231]

$$s_m(t+1) = \mu[1 - s_m(t)] + [1 - \Theta_m(\bar{\varrho})] s_m(t), \quad (5.4)$$

where  $\Theta_m(\bar{\varrho}) = 1 - [1 - \bar{\theta}(\bar{\varrho})]^m$  is the probability for a susceptible node of membership  $m$  to get infected.  $\bar{\varrho}(t)$  is the probability that a node belonging in any group is infected and  $\bar{\theta}(\bar{\varrho})$  is the probability for a susceptible node to get infected in any group,

$$\bar{\varrho}(t) = \sum_m [1 - s_m(t)] \frac{m g_m}{\langle m \rangle} \quad \text{and} \quad \bar{\theta}(\bar{\varrho}) = \sum_n \bar{\theta}_n(\bar{\varrho}) \frac{n p_n}{\langle n \rangle}, \quad (5.5)$$

where  $\bar{\theta}_n(\bar{\varrho})$  is the probability for a node to get infected in a group of size  $n$ . Because of the annealed structure,  $\bar{\theta}_n(\bar{\varrho})$  is just the average of  $\theta_n(\varrho)$  with  $\varrho = i/(n-1)$  over a binomial distribution,

$$\bar{\theta}_n(\bar{\varrho}) = \sum_{i=0}^{n-1} \binom{n-1}{i} \bar{\varrho}^i (1 - \bar{\varrho})^{n-1-i} \theta_n \left( \frac{i}{n-1} \right), \quad (5.6)$$

with  $\theta_n(\varrho)$  defined at Eq. (5.2).

Figure 5.2(b) shows a first consequence of the nonlinear kernel : The appearance of superexponential growth for the global prevalence  $I(t)$  when  $\nu > 1$ . Note that the growth is approximately exponential until a sufficiently high prevalence. For  $\nu \leq 1$ , we instead have a standard exponential growth until saturation is reached.

In the steady state of the epidemic dynamics, we obtain a self-consistent solution

$$s_m^* = \frac{\mu}{\mu + \Theta_m^*} \quad \text{and} \quad \bar{\varrho}^* = \sum_m \left( \frac{\Theta_m^*}{\mu + \Theta_m^*} \right) \frac{m g_m}{\langle m \rangle} \equiv G(\bar{\varrho}^*), \quad (5.7)$$

since  $\Theta_m$  is a function of  $\bar{\varrho}$ .

For contagions with a nonlinear infection kernel, the phase transition associated with the order parameter  $I^*$  can be continuous or discontinuous with a bistable regime. Consequently, we define the *invasion* threshold  $\beta_c$  such that for all  $\beta > \beta_c$ , the absorbing state  $I^* = 0$  is unstable [dashed line in Fig. 5.2(c)]. We also define the *persistence* threshold  $\beta_p$  such that for all  $\beta < \beta_p$ , the absorbing state  $I^* = 0$  is globally attractive [dotted line in Fig. 5.2(c)]. For continuous phase transitions,  $\beta_c$  and  $\beta_p$  coincide, and this is called the epidemic threshold; for a discontinuous phase transition,  $\beta_p < \beta_c$ , and for all  $\beta \in (\beta_p, \beta_c)$ , there exists typically three solutions,  $I_1^* = 0$  and  $I_2^*, I_3^* > 0$ , with  $I_1^*$  and  $I_3^*$  locally stable.

The invasion threshold  $\beta_c$  can be found by imposing  $G'(0) = 1$ . The persistence threshold  $\beta_p$  is obtained by imposing both  $\bar{\varrho}^* = G(\bar{\varrho}^*)$  and  $G'(\bar{\varrho}^*) = 1$  for  $\bar{\varrho}^* > 0$ , and any tricritical point can be found by imposing  $G'(0) = 1$  and  $G''(0) = 0$ .

In Appendix 5.9, we obtain an exact self-consistent expression for the invasion threshold, and using again an asymptotic approximation, we find

$$\beta_c \sim \left( \frac{\mu \langle n \rangle \langle m \rangle}{\langle n(n-1)^{1-\nu} [f(n)]^\nu \rangle \langle m^2 \rangle} \right)^{1/\nu}. \quad (5.8)$$

It depends in a intricate manner on both the moments of  $p_n$  and  $g_m$  and the nonlinearity of the infection kernel. As illustrated in Fig. 5.2(c), the invasion threshold can become very small for  $\nu < 1$  (vanishing for  $\nu \rightarrow 0$ ), even for homogeneous  $g_m$  and  $p_n$ . Note that to obtain a sublinear kernel  $\nu < 1$ , it typically requires a distribution  $\varphi(\tau) \propto \tau^{-\alpha-1}$  with  $\alpha < 1$ , which implies that the mean participation time  $\langle \tau \rangle$  diverges. In Appendix 5.9, we show that if instead we fix  $\langle \tau \rangle$  while varying  $\alpha$ , there exists an optimal temporal heterogeneity  $\alpha^*$  that minimizes the invasion threshold  $\beta_c$ , and maximizes early spread.

The minimal kernel exponent  $\nu_c$  leading to a discontinuous phase transition is given by a tricritical point (see Appendix 5.9). Although exact solutions require numerical evaluation, we get three insights from an asymptotic expansion.

- (i)  $\nu > 1$  is necessary in order to have a discontinuous phase transition, but it is not sufficient :  $\nu_c$  depends on the first three moments of  $g_m$  and, in a more complicated manner, on the distribution  $p_n$ .
- (ii) It is necessary to have environments of size  $n > 2$  to have a discontinuous phase transition.
- (iii) A more heterogeneous  $g_m$  leads to a larger  $\nu_c$ .

Similar observations were made in Ref. [170] for  $n \leq 3$ .

## 5.7 Conclusion

Our framework captures many properties usually overlooked for the sake of simplicity in epidemic models : the higher-order structure of contacts, the temporal heterogeneity of human activity, and the thresholding effects over the exposure due to the host immune system. In Appendix 5.8, we also demonstrate that our results are robust to variations in individual infectiousness or local transmission in different environments. In particular, we recover a universal nonlinear infection kernel that provides a connection between complex contagions based on nonlinear infection kernels [182] and threshold models [75, 112, 288].

Our results challenge a key assumption of most epidemic models and ask : Why assume a linear relationship between the number of infectious contacts and the risk of infection? This question is critical since three of the basic insights gathered from epidemic models break down under nonlinear infection kernels : They can lead to a discontinuous relationship between disease transmission and epidemic size, to a bistable regime where macroscopic outbreak and a disease-free state coexist, and to a superexponential growth. While the first two are difficult to assess for real contagions, superexponential spread has been observed for influenzalike illness [249].

Even though we considered the SIS model to simplify the analysis, the universal infection kernel  $\theta_n(\varrho) \sim \varrho^\nu$  could be directly integrated in more realistic models—such as SEIR or SIRS—where the same phenomenology typically carries over.

The phenomenology being drastically different from standard epidemiological models begs the following question : Why do linear models work? Even for a nonlinear kernel  $\theta_n(\varrho)$ , the probability of infection  $\bar{\theta}_n(\bar{\varrho})$  (averaged over group configurations) is *linear* in  $\bar{\varrho}$  if  $\bar{\varrho} \ll 1$  [see Eq. (5.6)]. Therefore, linear models are a good approximation when the prevalence is sufficiently low, but breaks down at higher prevalence, as clearly illustrated in Fig. 5.2(b) when  $\nu = 1.5$ .

The mathematical framework we use to solve the SIS model hinges on a mean-field or annealed approximation, as in other studies [88, 142, 147, 170], thereby suppressing *dynamical correlations* within groups. As we show in Appendix 5.10, dynamical correlations can be captured using approximate master equations [105, 125, 178, 189, 226, 266, 267], which are more complicated but provide similar results with better agreement to simulations. Future works could investigate more thoroughly the interplay between dynamical correlations, nonlinear kernels, and spatiotemporal heterogeneity.

Altogether, our conclusions stress the need to embrace heterogeneity in disease modeling—in the infection dynamics itself, in patterns of temporal activity, and in the higher-order structure of contact networks. Epidemics should be seen as the result of a collective process, where higher-order structure and temporal patterns can drive complex dynamics.

## 5.8 Appendix : Nonlinear infection kernel with general dose distribution

As in the main text, let us assume that a susceptible individual participates to an environment of size  $n$ , with a fraction  $\varrho$  of the  $n - 1$  individual being infected, for a duration  $\tau$ . During this period, the susceptible individual receives an infective dose  $\kappa$  of distribution  $\pi(\kappa; \lambda)$ , where  $\lambda \equiv \langle \kappa \rangle$  is

$$\lambda = \beta f(n) \tau \varrho . \tag{5.9}$$

The susceptible individual develops the infection if  $\kappa \geq K$ , hence with probability

$$\bar{\Pi}(K; \lambda) = \int_K^\infty \pi(\kappa; \lambda) d\kappa . \tag{5.10}$$

Note that the distribution could be discrete or continuous.

The infection kernel is then obtained by averaging over the participation time,

$$\theta_n(\varrho) = \int_{\mathcal{D}_\tau} \bar{\Pi}(K; \lambda) \wp(\tau) d\tau , \tag{5.11}$$

for some domain  $\mathcal{D}_\tau$ . In this chapter we focus on heterogeneous exposure modeled by  $\wp(\tau) = C_\alpha \tau^{-\alpha-1}$  with  $\tau \in [1, \tau_{\max}]$ .

As mentioned in the main text,  $\bar{\Pi}(K; \lambda)$  is ill-defined for  $\tau \geq \mathcal{T}$ , where  $\mathcal{T}$  is the clearing window for the immune system, but fortunately, we will be able to neglect these events. In fact, we always have that  $\bar{\Pi}(K; \lambda) \leq 1$ , hence the remainder associated to participation times  $\tau \geq \mathcal{T}$  is upper-bounded

$$\mathcal{R}(\mathcal{T}) \leq \int_{\mathcal{T}}^{\tau_{\max}} \wp(\tau) d\tau = C_\alpha \int_{\mathcal{T}}^{\tau_{\max}} \tau^{-\alpha-1} d\tau = O(\mathcal{T}^{-\alpha}), \quad (5.12)$$

As it will become clear in this section, terms  $O(\mathcal{T}^{-\alpha})$  have a negligible contribution to the infection kernel. For mathematical convenience, we thus redefine our support as  $\tau \in [1, \mathcal{T}]$  and the infection kernel as

$$\theta_n(\varrho) = C_\alpha \int_1^{\mathcal{T}} \bar{\Pi}(K; K\tau/\tau_c) \tau^{-\alpha-1} d\tau, \quad (5.13)$$

where, as in the main text,  $\tau_c \equiv K/\beta f(n)\varrho$  is the characteristic time for an individual to become infected. Integrating by parts, we have

$$\theta_n(\varrho) = \frac{C_\alpha}{\alpha} \left[ \bar{\Pi}(K; K/\tau_c) - \bar{\Pi}(K; K\mathcal{T}/\tau_c) \mathcal{T}^{-\alpha} + \int_1^{\mathcal{T}} \partial_\tau \bar{\Pi}(K; K\tau/\tau_c) \tau^{-\alpha} d\tau \right]. \quad (5.14)$$

Making the change of variable  $x = \tau/\tau_c$ , we obtain

$$\theta_n(\varrho) = \frac{C_\alpha}{\alpha} \left[ \bar{\Pi}(K; K/\tau_c) - \bar{\Pi}(K; K\mathcal{T}/\tau_c) \mathcal{T}^{-\alpha} + \tau_c^{-\alpha} \int_{1/\tau_c}^{\mathcal{T}/\tau_c} \partial_x \bar{\Pi}(K; Kx) x^{-\alpha} dx \right]. \quad (5.15)$$

As we will make clear with a few cases, in the regime where  $1 \ll \tau_c \ll \mathcal{T}$ , depending on the asymptotic properties of each term in Eq. (5.15), many distributions  $\pi(\kappa; \lambda)$  map to the universal nonlinear infection kernel

$$\theta_n(\varrho) \simeq D_\alpha \tau_c^{-\nu} \propto \varrho^\nu, \quad (5.16)$$

for some proportionality constant  $D_\alpha$ .

### 5.8.1 Dose distribution where $\lambda$ is a scale parameter

An important subclass is all distributions  $\pi(\kappa; \lambda)$  where  $\lambda$  is a *scale parameter*, i.e.,

$$\begin{aligned} \pi(\kappa; \lambda) &= \frac{1}{\lambda} \pi\left(\frac{\kappa}{\lambda}; 1\right) \equiv \frac{1}{\lambda} \pi\left(\frac{\kappa}{\lambda}\right), \\ \bar{\Pi}(K; \lambda) &= \bar{\Pi}\left(\frac{K}{\lambda}; 1\right) \equiv \bar{\Pi}\left(\frac{K}{\lambda}\right). \end{aligned}$$

In this case, Eq. (5.15) becomes

$$\theta_n(\varrho) = \frac{C_\alpha}{\alpha} \left[ \bar{\Pi}(\tau_c) - \bar{\Pi}(\tau_c/\mathcal{T}) \mathcal{T}^{-\alpha} + \tau_c^{-\alpha} \int_{1/\tau_c}^{\mathcal{T}/\tau_c} \partial_x \bar{\Pi}(1/x) x^{-\alpha} dx \right]. \quad (5.17)$$

Making the change of variable  $y = 1/x$ , and recall that  $\bar{\Pi}$  is the *complementary* cumulative distribution, we obtain

$$\theta_n(\varrho) = \frac{C_\alpha}{\alpha} \left[ \bar{\Pi}(\tau_c) - \bar{\Pi}(\tau_c/\mathcal{T}) \mathcal{T}^{-\alpha} + \tau_c^{-\alpha} \int_{\tau_c/\mathcal{T}}^{\tau_c} \pi(y) y^\alpha dy \right]. \quad (5.18)$$

In the regime  $1 \ll \tau_c \ll \mathcal{T}$ , the integral on the right is bounded from below by some constant, or in asymptotic notation it is  $\Omega(1)$ , hence the last term is  $\Omega(\tau_c^{-\alpha})$ , whereas the second term is  $O(\mathcal{T}^{-\alpha})$  so we can neglect it. This also justifies neglecting participation times  $\tau \geq \mathcal{T}$  since  $\mathcal{R}(\mathcal{T}) = O(\mathcal{T}^{-\alpha})$ . The infection kernel becomes

$$\theta_n(\varrho) \simeq \frac{C_\alpha}{\alpha} \left[ \bar{\Pi}(\tau_c) + \tau_c^{-\alpha} \int_{\tau_c/\mathcal{T}}^{\tau_c} \pi(y)y^\alpha dy \right]. \quad (5.19)$$

It is now clear that for all distributions  $\pi(y)$  that decrease faster than  $\tau_c^{-\alpha-1}$  in the tail, the integral converges to a constant  $D_\alpha$  and  $\bar{\Pi}(\tau_c) \ll \tau_c^{-\alpha}$ , therefore, we recover the infection kernel

$$\theta_n(\varrho) \simeq D_\alpha \tau_c^{-\alpha} \propto \varrho^\alpha. \quad (5.20)$$

If instead  $\pi(y)$  decreases slower than  $y^{-\alpha-1}$ , let's say as  $y^{-\xi-1}$  for large  $y$ , where  $\xi < \alpha$ , then both terms are asymptotically proportional to  $\tau_c^{-\xi}$ ,

$$\theta_n(\varrho) \simeq E_\alpha \tau_c^{-\xi} \propto \varrho^\xi, \quad (5.21)$$

where  $E_\alpha$  is another constant prefactor. If  $\xi = \alpha$ , we would have

$$\theta_n(\varrho) \sim \varrho^\alpha \ln(1/\varrho). \quad (5.22)$$

So except for this limit case, we always have

$$\theta_n(\varrho) \sim \varrho^\nu, \quad (5.23)$$

with  $\nu = \min(\alpha, \xi)$ . This is illustrated in Fig. 5.3 with the Weibull and the Fréchet (inverse Weibull) distributions.

### 5.8.2 Other cases

When  $\lambda$  is a scale parameter, the properties of the dose distribution are constrained, for instance the fact that  $\langle \kappa \rangle = \lambda$ , this implies that  $\text{Var}(\kappa) = \langle \kappa^2 \rangle - \langle \kappa \rangle^2 \propto \lambda^2$ . Here we investigate other scenarios.

Perhaps the simplest case is to consider  $\pi(\kappa; \lambda) = \delta(\kappa - \lambda)$ , i.e., the dose is no longer a random variable, it is strictly proportional to the time spent in the environment and to the number of infected people. In this case,  $\text{Var}(\kappa) = 0$ . We directly have

$$\bar{\Pi}(K; K\mathcal{T}/\tau_c) = \vartheta(K\mathcal{T}/\tau_c - K) = \vartheta(\mathcal{T} - \tau_c), \quad (5.24)$$

where  $\vartheta(\dots)$  is the Heaviside function, and therefore

$$\theta_n(\varrho) = C_\alpha \int_1^{\mathcal{T}} \vartheta(\mathcal{T} - \tau_c) \tau^{-\alpha-1} d\tau, \quad (5.25)$$

$$= C_\alpha \int_{\tau_c}^{\mathcal{T}} \tau^{-\alpha-1} d\tau \quad \text{if } 1 \leq \tau_c, \quad (5.26)$$

$$\simeq \frac{C_\alpha}{\alpha} \tau_c^{-\alpha} \quad \text{if } \tau_c \ll \mathcal{T}, \quad (5.27)$$

$$\propto \varrho^\alpha. \quad (5.28)$$

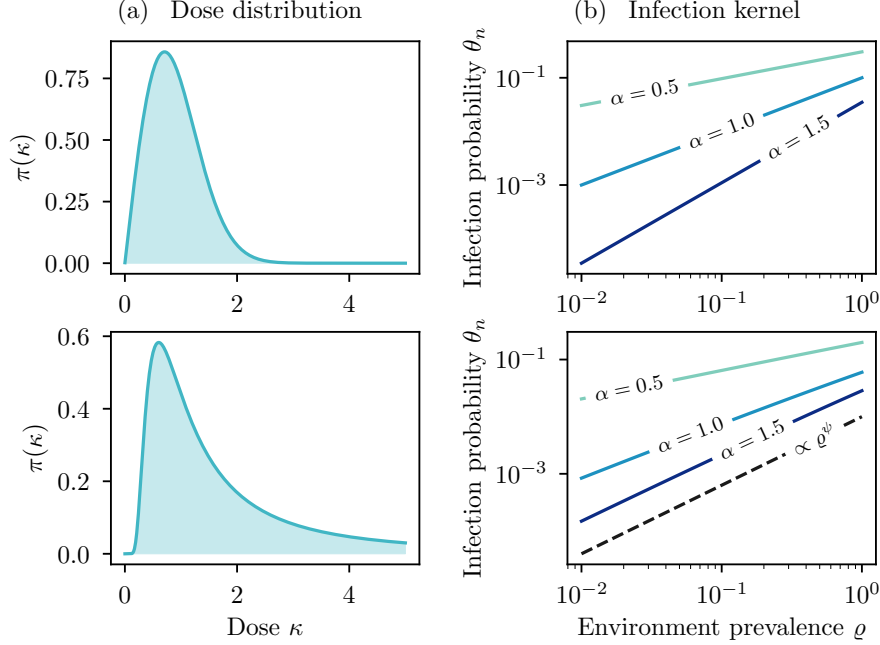


FIGURE 5.3 – Infection kernel resulting from different dose distributions, when considering  $\lambda$  to be scale parameter, i.e.,  $\sigma \propto \lambda$ . (a) Dose distributions, respectively the Weibull  $\pi(\kappa) \propto x^{\psi-1}e^{-(\kappa/\sigma)^\psi}$  and the Fréchet  $\pi(\kappa) \propto x^{-\psi-1}e^{-(\kappa/\sigma)^{-\psi}}$  distribution from top to bottom. We use  $\psi = 2$  for the Weibull and  $\psi = 1.2$  for the Fréchet, and fix  $\sigma = 1$  (on the left) for visual purpose only. (b) We fix  $f(n) = 1$ ,  $\beta = 0.1$ ,  $\mathcal{T} \rightarrow \infty$ , and ensure  $\langle \kappa \rangle = \lambda$  by using  $\sigma \propto \lambda$ . For the Fréchet distribution,  $\nu = \min(\alpha, \psi)$ , hence for  $\alpha = 1.5$ , we see that the kernel exponent is  $\psi = 1.2$ .

We thus recover the same result as before under the less restrictive condition  $1 \leq \tau_c \ll \mathcal{T}$ . This suggests that if  $\pi(\kappa; \lambda)$  is sufficiently concentrated around its mean, our result should hold.

Another case study we consider is when  $\text{Var}(\kappa) \propto \lambda^{2-z}$ , where  $z \in [0, 1]$ . We achieve this using the gamma distribution

$$\pi(\kappa; \sigma, \psi) = \frac{\kappa^{\psi-1} e^{-\kappa/\sigma}}{\Gamma(\psi) \sigma^\psi}. \quad (5.29)$$

Since the mean of the distribution is  $\langle \kappa \rangle = \psi\sigma$  and the variance is  $\text{Var}(\kappa) = \psi\sigma^2$ , we can set  $\sigma = \lambda^{1-z}$  and  $\psi = \lambda^z$ , in which case we respect  $\langle \kappa \rangle = \lambda$  and have  $\text{Var}(\kappa) = \lambda^{2-z}$ . We are thus able to interpolate between two cases :  $\lambda$  being a scale parameter ( $z = 0$ ), for which we know our result hold, and where the mean and the variance are the same ( $z = 1$ ). We do not provide an analytical proof, but we show in Fig. 5.4 that the kernel  $\theta_n \sim \rho^\alpha$  appears to hold for all  $z < 1$ , whereas for  $z = 1$ , we have a kernel  $\theta_n \sim \rho^\nu$  where  $\nu = \min(\alpha, 1)$ . These again belong to the general class of universal power-law infection kernels.

A last case we consider is when  $\pi(\kappa; \lambda)$  is discrete, i.e.,  $\kappa \in \mathbb{N}_0$ . Let us consider more specifically the Poisson distribution

$$\pi(\kappa; \lambda) = \frac{\lambda^\kappa e^{-\lambda}}{\kappa!}. \quad (5.30)$$

This distribution can be justified from a mechanistic perspective : let us assume that individuals from a group interact at a constant rate  $\beta f(n)$ . During the period of duration  $\tau$ , a susceptible individual



has on average  $\lambda \equiv \beta f(n) \rho \tau$  interactions with infected individuals (doses), and the distribution for  $\kappa$  is the Poisson distribution above.

Without loss of generality, we assume that  $K \in \mathbb{N}$  is a small positive integer. It will be more convenient to work with  $\tilde{\tau}_c \equiv 1/\lambda f(n)\beta = \tau_c/K$ , the characteristic time to receive a unit dose. We rewrite Eq. (5.15) as

$$\theta_n(\varrho) = \frac{C_\alpha}{\alpha} \left[ \bar{\Pi}(K; 1/\tilde{\tau}_c) - \bar{\Pi}(K; \mathcal{T}/\tilde{\tau}_c) \mathcal{T}^{-\alpha} + \tilde{\tau}_c^{-\alpha} \int_{1/\tilde{\tau}_c}^{\mathcal{T}/\tilde{\tau}_c} \partial_x \bar{\Pi}(K; x) x^{-\alpha} dx \right]. \quad (5.31)$$

We have that the complementary cumulative distribution is

$$\bar{\Pi}(K; x) = 1 - \sum_{\kappa=0}^{K-1} \frac{x^\kappa e^{-x}}{\kappa!}, \quad (5.32)$$

and

$$\partial_x \bar{\Pi}(K; x) = - \sum_{\kappa=0}^{K-1} \frac{1}{\kappa!} \frac{d}{dx} x^\kappa e^{-x} = e^{-x} + \sum_{\kappa=1}^{K-1} \frac{1}{\kappa!} (x^\kappa e^{-x} - \kappa x^{\kappa-1} e^{-x}) \quad (5.33)$$

For the sake of simplicity, let us consider  $K = 2$ , which implies

$$\bar{\Pi}(K = 2; 1/\tilde{\tau}_c) = 1 - (1 + \tilde{\tau}_c^{-1}) e^{-1/\tilde{\tau}_c}, \quad (5.34)$$

$$\simeq \frac{\tilde{\tau}_c^{-2}}{2} \quad \text{for } \tau_c \gg 1, \quad (5.35)$$

$$\bar{\Pi}(K = 2; \mathcal{T}/\tilde{\tau}_c) \simeq 1 \quad \text{for } \mathcal{T} \gg \tau_c, \quad (5.36)$$

and

$$\partial_x \bar{\Pi}(K = 2; x) = x e^{-x}. \quad (5.37)$$

Inserting these in Eq. (5.31), we obtain

$$\theta_n(\varrho) \simeq \frac{C_\alpha}{\alpha} \left[ \frac{\tilde{\tau}_c^{-2}}{2} + \mathcal{T}^{-\alpha} + \tilde{\tau}_c^{-\alpha} \int_{1/\tilde{\tau}_c}^{\mathcal{T}/\tilde{\tau}_c} x^{1-\alpha} e^{-x} dx \right]. \quad (5.38)$$

The integral converges to  $\Gamma(2 - \alpha)$  if  $1 \ll \tilde{\tau}_c \ll \mathcal{T}$  and  $\alpha < 2$ . If  $\alpha = 2$ , the integral diverges as  $\ln(\tilde{\tau}_c)$  and for  $\alpha > 2$ , it diverges as  $\tilde{\tau}_c^{\alpha-2}$ . Consequently, we have

$$\theta_n(\varrho) \sim \begin{cases} \varrho^\alpha & \text{if } \alpha < 2, \\ \varrho^2 \ln(1/\varrho) & \text{if } \alpha = 2, \\ \varrho^2 & \text{if } \alpha > 2. \end{cases} \quad (5.39)$$

Ignoring the limit case  $\alpha = 2$ , we have  $\theta_n(\varrho) \sim \varrho^\nu$  where  $\nu = \min(\alpha, 2)$ . In the general case  $K \neq 2$ , similarly, we can show that  $\nu = \min(\alpha, K)$ .

### 5.8.3 Beyond a pure power-law distribution for participation time

Our previous results assume that  $\varphi(\tau)$  is a pure power-law distribution. However, we can relax this assumption by assuming that  $\varphi(\tau) \propto \tau^{-\alpha}$  for  $\tau > \tau^*$ . In this case, it is straightforward to generalize

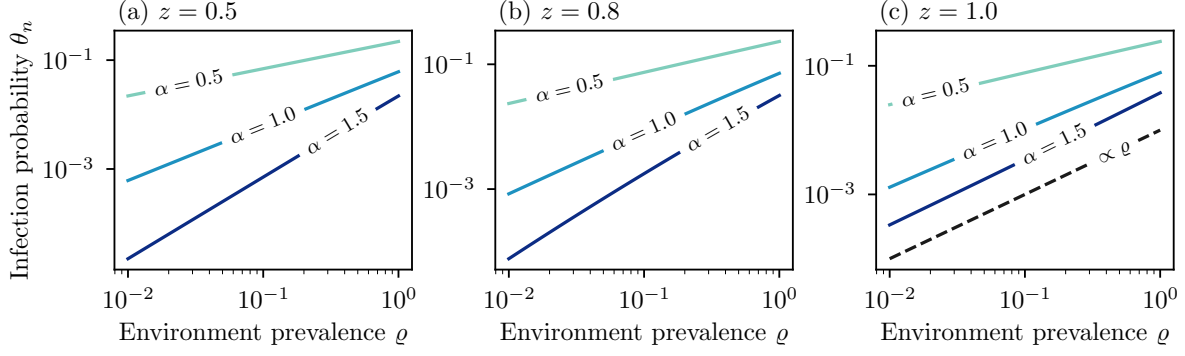


FIGURE 5.4 – Infection kernel when  $\lambda$  is no longer a scale parameter. We use the gamma distribution at Eq. (5.29) with  $\sigma = \lambda^{1-z}$  and  $\psi = \lambda^z$ , with different values of  $z$ . We use  $f(n) = 1$ ,  $\beta = 0.1$  and  $\mathcal{T} \rightarrow \infty$ .

our results by splitting the domain  $[1, \mathcal{T}]$  in two parts,  $[1, \tau^*]$  and  $[\tau^*, \mathcal{T}]$ . Equation (5.13) becomes

$$\theta_n(\varrho) = \int_1^{\tau^*} \bar{\Pi}(K; K\tau/\tau_c) \wp(\tau) d\tau + C_\alpha \int_{\tau^*}^{\mathcal{T}} \bar{\Pi}(K; K\tau/\tau_c) \tau^{-\alpha-1} d\tau, \quad (5.40)$$

where  $C_\alpha$  is again a normalization factor. Note that the first term can be upper-bounded,

$$\int_1^{\tau^*} \bar{\Pi}(K; K\tau/\tau_c) \wp(\tau) d\tau \leq (\tau^* - 1) M \bar{\Pi}(K; K\tau^*/\tau_c), \quad (5.41)$$

where  $M = \max\{\wp(\tau) | \tau \in [1, \tau^*]\}$  is some constant. Consequently, Eq. (5.40) becomes

$$\theta_n(\varrho) = O[\bar{\Pi}(K; K\tau^*/\tau_c)] + C_\alpha \int_{\tau^*}^{\mathcal{T}} \bar{\Pi}(K; K\tau/\tau_c) \tau^{-\alpha-1} d\tau, \quad (5.42)$$

Integrating by parts the second term, we obtain

$$\theta_n(\varrho) = \frac{C_\alpha}{\alpha} \left[ O[\bar{\Pi}(K; K\tau^*/\tau_c)] - \bar{\Pi}(K; K\mathcal{T}/\tau_c) \mathcal{T}^{-\alpha} + \tau_c^{-\alpha} \int_{\tau^*/\tau_c}^{\mathcal{T}/\tau_c} \partial_x \bar{\Pi}(K; Kx) x^{-\alpha} dx \right]. \quad (5.43)$$

Therefore, noting the similarity with Eq. (5.15), all our previous results hold in the more restrictive regime  $\tau^* \ll \tau_c \ll \mathcal{T}$ .

Instead of having a piecewise distribution  $\wp(\tau)$ , many real distributions are asymptotically power law, i.e.,  $\wp(\tau) \propto \tau^{-\alpha-1}$  for sufficiently large  $\tau$ . One just have to pick a  $\tau^*$  large enough such that  $\wp(\tau)$  is a sufficiently good approximation of a power law for  $\tau > \tau^*$ .

#### 5.8.4 Breaking the conditions of the asymptotic approximation

For the asymptotic developments above, we assume that  $1 \ll \tau_c \ll \mathcal{T}$ . In Fig. 5.5, we illustrate what happens when this is not respected. For the sake of simplicity, let us say that the conditions are broken if  $\tau_c \leq 10$  or  $\mathcal{T}/\tau_c \leq 10$ .

Let us recall that  $\tau_c = 1/\beta f(n)\varrho$ . In Fig. 5.5(a), we pick  $\beta$  and  $f(n)$  such that  $\tau_c \leq 10$  for  $\varrho \geq 0.1$  (on the right of the vertical dashed line), in which case the condition  $\tau_c \gg 1$  is not respected. We see that the scaling  $\varrho^\alpha$  is lost in this regime, but is still valid for  $\varrho \lesssim 0.1$ . In Fig. 5.5(b), we pick  $\mathcal{T}$ ,  $\beta$  and  $f(n)$

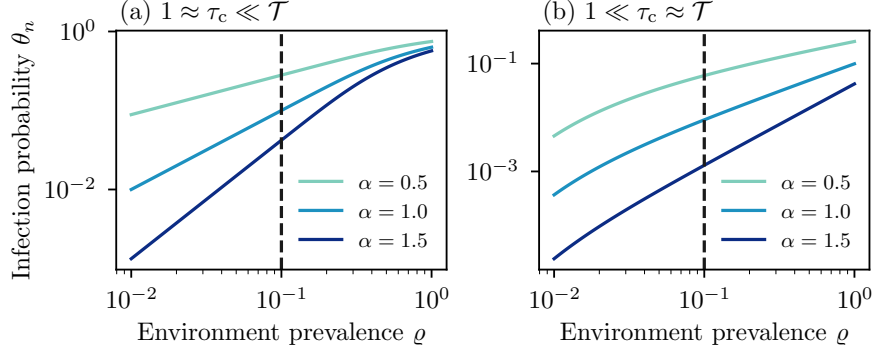


FIGURE 5.5 – Infection kernel when the condition  $1 \ll \tau_c \ll \mathcal{T}$  is no longer respected. We use an exponential dose distribution  $\pi(\kappa; \lambda) = e^{\kappa/\lambda}$  and  $f(n) = 1$ . (a)  $\beta = 1$  and  $\mathcal{T} \rightarrow \infty$ . The vertical dashed line indicates when  $\tau_c = 10$ . (b)  $\beta = 0.1$  and  $\mathcal{T} = 10^3$ . The vertical dashed line indicates when  $\tau_c = 100 = 0.1\mathcal{T}$ . Note that  $\tau_c$  increases when  $\rho$  decreases.

such that  $\mathcal{T}/\tau_c \leq 10$  for  $\rho \leq 0.1$  (on the left of the vertical dashed line), in which case the condition  $\mathcal{T} \gg \tau_c$  is not respected. We see that the scaling  $\rho^\alpha$  is lost in this regime, but is still valid for  $\rho \gtrsim 0.1$ .

In summary, the condition  $\mathcal{T} \gg \tau_c$ , for which we can choose a specific tolerance such as  $\mathcal{T}/\tau_c \geq 10$ , gives a lower bound  $\rho_1$  for the validity of  $\theta_n(\rho) \sim \rho^\alpha$ , while the condition  $\tau_c \gg 1$ , or as we chose  $\tau_c \geq 10$ , gives an upper bound  $\rho_2$ . We can therefore assess that the infection kernel is valid on some domain  $[\rho_1, \rho_2]$ .

### 5.8.5 Robustness to other sources of heterogeneity

The results put forward in this chapter highlight the role of the burstiness of interactions for the emergence of nonlinear infection kernels. The heterogeneity of membership  $m$  and environment sizes  $n$  is also incorporated in the modeling approach we use in the main text. In this section, we show that adding other layers of heterogeneity do not alter the form of the universal kernel  $\theta_n(\rho) \sim \rho^\nu$ . We demonstrate it in the case where  $\lambda$  is a scaling parameter for the sake of simplicity.

#### Heterogeneity of environments

The attack rate in an environment is certainly affected by the conditions of the environment itself : for instance, nonventilated places might be more prone to infections than better ventilated ones. Let us assume that  $\beta \in [0, \infty)$  are i.i.d random variables for each environment, drawn from an arbitrary distribution  $Q(\beta)$ .

If  $Q(\beta)$  is a power law, it plays a similar role as  $\wp(\tau)$ , since  $\beta$  and  $\tau$  appear side by side in the definition of the mean dose  $\langle \kappa \rangle = \lambda$ . With  $\tau$  and  $\beta$  both distributed as power laws, the distribution  $R(x)$  for  $x = \tau\beta$  is a mixture of power-law distributions [214]. The tail of  $R(x)$  is then dominated by the slowest decaying distribution, either  $Q(\beta)$  or  $\wp(\tau)$ . In all cases, we could extend the analysis above using  $R(x)$  instead  $\wp(\tau)$ , and we would recover that  $\theta_n(\rho) \sim \rho^\nu$  with  $\nu$  being the exponent of the distribution that decreases the slowest (see Fig. 5.6 lower row). We therefore assume that  $Q(\beta)$  decreases faster than a power law, i.e.,  $Q(\beta) \ll \beta^{-a} \forall a \in \mathbb{N}$  in the limit  $\beta \rightarrow \infty$ .

Recall that  $\tau_c^{-1} \equiv f(n)\beta\varrho$ , hence  $\tau_c = \tau_c(\beta)$ . We can rewrite Eq. (5.18) as

$$\theta_n(\varrho) = \int_0^\infty \frac{C_\alpha}{\alpha} \left[ \bar{\Pi}(\tau_c) - \bar{\Pi}(\tau_c/\mathcal{T})\mathcal{T}^{-\alpha} + \tau_c^{-\alpha} \int_{\tau_c/\mathcal{T}}^{\tau_c} \pi(y)y^\alpha dy \right] Q(\beta)d\beta . \quad (5.44)$$

Let now us suppose that we can divide the support for  $\beta$  in three subintervals,  $\mathcal{B}_1 \equiv [0, \beta_1]$ ,  $\mathcal{B}_2 \equiv [\beta_1, \beta_2]$  and  $\mathcal{B}_3 \equiv [\beta_2, \infty)$ , with  $\int_{\beta \in \mathcal{B}_2} Q(\beta)d\beta$  sufficiently large. In the middle subinterval, we assume that the usual condition for asymptotic expansion holds,  $1 \ll \tau_c(\beta) \ll \mathcal{T}$  for all  $\beta \in \mathcal{B}_2$ . In the other subintervals, this condition is not respected for all  $\beta$ , i.e., there are values for  $\beta \in \mathcal{B}_1$  such that  $\tau_c(\beta) \approx \mathcal{T}$  or even  $\tau_c(\beta) \gg \mathcal{T}$  and values for  $\beta \in \mathcal{B}_2$  such that  $\tau_c(\beta) \approx 1$  or even  $\tau_c(\beta) \ll 1$ . We investigate the asymptotic behavior of  $\theta_n^{\mathcal{B}_j}$  for each subinterval  $\mathcal{B}_j$ , where  $\theta_n = \sum_{j=1}^3 \theta_n^{\mathcal{B}_j}$ .

In the subinterval  $\mathcal{B}_2$ , we can replace the inside of the integral on  $\beta$  with the usual asymptotic expansion, leading to

$$\theta_n^{\mathcal{B}_2}(\varrho) \simeq D_\alpha \int_{\beta \in \mathcal{B}_2} \tau_c^{-\alpha}(\beta) Q(\beta)d\beta , \quad (5.45)$$

$$= D_\alpha \varrho^\alpha f(n)^\alpha K^{-\alpha} \int_{\beta \in \mathcal{B}_2} \beta^\alpha Q(\beta)d\beta , \quad (5.46)$$

$$\propto \varrho^\alpha . \quad (5.47)$$

In the subinterval  $\mathcal{B}_1$ , note that smaller values of  $\tau_c$  increases the value of the integrand on  $\beta$ —by definition,  $\lambda \propto \tau_c^{-1}$ , hence the mean dose received is larger for smaller  $\tau_c$ . Therefore, we can bound from above  $\theta_n^{\mathcal{B}_1}$  by using for all  $\beta$  the smallest value possible on the subinterval,  $\tau_c(\beta) \mapsto \tau_c(\beta_1) \ll \mathcal{T}$ . In this case, the asymptotic expansion we used for  $\mathcal{B}_2$  is again possible, which implies

$$\theta_n^{\mathcal{B}_1}(\varrho) = O[\tau_c^{-\alpha}(\beta_1)] . \quad (5.48)$$

However, Eq. (5.45) implies that  $\theta_n^{\mathcal{B}_2}(\varrho) = \Omega[\tau_c^{-\alpha}(\beta_1)]$ , which suggest that we can have  $\theta_n^{\mathcal{B}_1}(\varrho) \ll \theta_n^{\mathcal{B}_2}(\varrho)$ . In fact, this is respected if

$$\beta_1^\alpha \int_{\beta \in \mathcal{B}_1} Q(\beta)d\beta \ll \int_{\beta \in \mathcal{B}_2} \beta^\alpha Q(\beta)d\beta , \quad (5.49)$$

in which case we can neglect the contribution from  $\theta_n^{\mathcal{B}_1}(\varrho)$

Finally, in the subinterval  $\mathcal{B}_3$ , we again use the fact that smaller values of  $\tau_c$  increases the value of the integrand on  $\beta$ . In the limit  $\beta \rightarrow \infty$ ,  $\tau_c \rightarrow 0$ , and we are left with the upper bound

$$\theta_n^{\mathcal{B}_3}(\varrho) \leq \int_{\beta \in \mathcal{B}_3} Q(\beta)d\beta . \quad (5.50)$$

For the universal kernel  $\theta_n(\varrho) \sim \varrho^\alpha$  to hold, we must have

$$\theta_n^{\mathcal{B}_3}(\varrho) \ll \theta_n^{\mathcal{B}_2}(\varrho) . \quad (5.51)$$

A sufficient condition is if

$$\int_{\beta \in \mathcal{B}_3} Q(\beta)d\beta \ll D_\alpha \varrho^\alpha f(n)^\alpha K^{-\alpha} \int_{\beta \in \mathcal{B}_2} \beta^\alpha Q(\beta)d\beta . \quad (5.52)$$

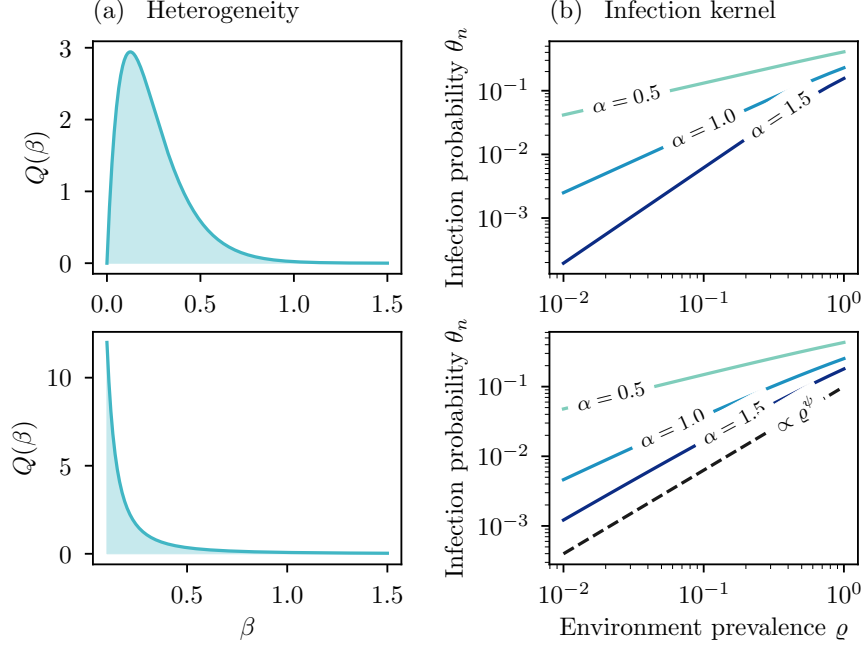


FIGURE 5.6 – Nonlinear infection kernels are robust to heterogeneous infectiousness. (a) Distributions for the heterogeneity of the environment. On the upper row, we use a gamma distribution  $Q(\beta) \propto \beta^{\psi-1}e^{-\beta/\sigma}$  with  $\sigma = 1/8$ ,  $\psi = 2$ , and  $\beta \in [0, \infty)$ . On the lower row, we use a power-law distribution  $Q(\beta) \propto \beta^{-\psi}$  with  $\psi = 1.2$  and  $\beta \in [0.5, \infty)$ . (b) Infection kernel obtained using the distribution  $Q(\beta)$  on the left. We use  $\pi(\kappa; \lambda) \propto e^{-\kappa/\lambda}$ ,  $\wp(\tau) \propto \tau^{-\alpha-1}$ ,  $\mathcal{T} \rightarrow \infty$  and  $f(n) = 0.1$ . For the gamma distribution (upper row), we have  $\theta_n(\varrho) \sim \varrho^\alpha$ , while for the power-law distribution (lower row), we have  $\theta_n(\varrho) \sim \varrho^\nu$  with  $\nu = \min(\alpha, \psi)$  and  $\psi = 1.2$ .

More intuitively, conditions at Eqs (5.49) and (5.52) indicate that the probability for  $\beta \in \mathcal{B}_1$  and  $\beta \in \mathcal{B}_3$  must be sufficiently small, i.e., environments of low or intense spreading are rare compared with environments where the spreading is moderate, satisfying  $1 \ll \tau_c(\beta) \ll \mathcal{T}$ . Infection kernels obtained with different  $Q(\beta)$  are presented in Fig. 5.6, where we once again recover the form  $\theta_n(\varrho) \sim \varrho^\nu$ .

### Heterogeneity of individual infectiousness

Similarly, all infectious individuals are not equivalent. For instance, the viral load within an infectious individual depends on the time since infection. Or more simply, people's behavior varies widely, some follow diligently public health recommendations (hand-washing, social distancing, etc.) and some do not. This motivates assigning i.i.d. random variables  $\beta_j$  with distribution  $\tilde{Q}(\beta_j)$  to characterize the infectiousness of each infected individual  $j \in \{1, \dots, i\}$ , where  $i \equiv \varrho(n-1)$ . A susceptible individual then receives a dose  $\kappa_j$  with distribution  $\tilde{\pi}(\kappa_j; \lambda_j)$  from each infectious individual  $j$ , where  $\langle \kappa_j \rangle = \lambda_j = \beta_j f(n) \tau \varrho$ .

However, it is straightforward to verify that adding this additional source of heterogeneity does not affect our results either. If we define  $\beta = \sum_{j=1}^i \beta_j$ , with distribution  $Q(\beta)$ , and  $\lambda = \sum_{j=1}^i \lambda_j$ , then  $\kappa = \sum_{j=1}^i \kappa_j$  has a distribution  $\pi(\kappa; \lambda)$  corresponding to

$$\pi(\kappa; \lambda) = \frac{1}{Q(\beta)} \int_0^\infty \cdots \int_0^\infty \delta\left(\kappa - \sum_{j=1}^i \kappa_j\right) \delta\left(\beta - \sum_{j=1}^i \beta_j\right) \prod_{j=1}^i \left\{ \tilde{\pi}(\kappa_j; \lambda_j) \tilde{Q}(\beta_j) d\kappa_j d\beta_j \right\}. \quad (5.53)$$

Therefore, this is totally equivalent to the case of a heterogeneous environment, with the infection kernel  $\theta_n(\varrho)$  defined by Eq. (5.44).

### Binomial distribution for the participation to groups

In the main text, it is assumed that a node of membership  $m$  interacts in  $m$  different environments *at each time step*. More realistically, we can consider that an individual participate to a group with probability  $q$ , hence the effective membership of a node for a time window is drawn from a binomial distribution with parameters  $m$  and  $q$ .

This can be incorporated in our approach simply by using

$$\wp(\tau) = qC_\alpha\tau^{-\alpha-1} + (1-q)\delta(\tau). \quad (5.54)$$

The effect is simply to map  $\theta_n(\varrho) \mapsto q\theta_n(\varrho)$ , leaving its scaling with  $\varrho$  unaltered. Therefore, the whole phenomenology is preserved.

### 5.8.6 More general temporal patterns

For a specific environment, we suggest in the main text that the participation time  $\tau_j$  of each individual  $j$  is i.i.d., with distribution  $\wp(\tau)$ . However, our result about the infection kernel does not require the independence of the participation time. For instance, the participation time of all individuals belonging to an environment could be perfectly correlated, i.e.,  $\tau_j = \tau$  for each individual  $j$  belonging to the environment. This could be more representative of certain scenarios, for example, of certain workplaces where people have the exact same schedule. More generally, we expect some varying level of correlation for the participation time to different types of environments.

The point is that correlations would be reflected on the form of the dose distribution  $\pi(\kappa; \lambda)$  and its mean—a more correlated participation time would be associated with a larger rate of dose accumulation  $\beta$  since the individuals spend more time altogether in the environment. Since our result does not depend on the form of  $\pi(\kappa; \lambda)$ , if the condition  $1 \ll \tau_c \ll \mathcal{T}$  is respected, so should be our result about the infection kernel  $\theta_n(\varrho) \sim \varrho^\nu$ .

## 5.9 Appendix : Characterization of the phase transition

As discussed in the main text, the phase transition is characterized by  $G(\bar{\varrho}^*)$ . In this section, we derive self-consistent equations for both the invasion threshold  $\beta_c$  and any tricritical point  $(\beta_c, \nu_c)$ , where  $\nu_c$  is a limit value for the nonlinear kernel exponent separating a continuous and a discontinuous phase transition.

All time-varying quantities are assumed to have reached the steady state—to simplify the notation, we drop the asterisk.

First, let us rewrite  $G(\bar{\varrho})$  [Eq. (5.7) in the main text] and its derivatives as

$$\begin{aligned} G(\bar{\varrho}) &= \frac{1}{\langle m \rangle} \left\langle \frac{m\Theta_m}{\mu + \Theta_m} \right\rangle, \\ G'(\bar{\varrho}) &= \frac{1}{\langle m \rangle} \left\langle \frac{m\Theta'_m(\bar{\varrho})}{\mu + \Theta_m} - \frac{m\Theta_m\Theta'_m(\bar{\varrho})}{(\mu + \Theta_m)^2} \right\rangle, \\ G''(\bar{\varrho}) &= \frac{1}{\langle m \rangle} \left\langle \frac{m\Theta''_m(\bar{\varrho})}{\mu + \Theta_m} - \frac{2m[\Theta'_m(\bar{\varrho})]^2}{(\mu + \Theta_m)^2} - \frac{m\Theta_m\Theta''_m(\bar{\varrho})}{(\mu + \Theta_m)^2} + \frac{2m\Theta_m[\Theta'_m(\bar{\varrho})]^2}{(\mu + \Theta_m)^3} \right\rangle. \end{aligned}$$

In the limit  $\bar{\varrho} \rightarrow 0$ , we have  $\bar{\theta} \rightarrow 0$ , which implies  $\Theta_m \rightarrow 0$  for all  $m$  and

$$\begin{aligned} \Theta'_m(0) &= m\bar{\theta}'(0), \\ \Theta''_m(0) &= m\bar{\theta}''(0) - m(m-1)[\bar{\theta}'(0)]^2. \end{aligned}$$

Therefore, we have

$$G'(0) = \frac{\langle m \rangle}{\mu} \bar{\theta}'(0), \quad (5.55a)$$

$$G''(0) = \frac{1}{\mu} \left\{ \langle m^2 \rangle \bar{\theta}''(0) - \left( \langle m^2(m-1) \rangle + \frac{2\langle m^3 \rangle}{\mu} \right) [\bar{\theta}'(0)]^2 \right\}. \quad (5.55b)$$

Second,  $\bar{\theta}(\bar{\varrho})$  is rewritten as

$$\bar{\theta}(\bar{\varrho}) = \frac{1}{\langle n \rangle} \langle n\bar{\theta}_n(\bar{\varrho}) \rangle, \quad (5.56)$$

hence it depends on  $\bar{\varrho}$  only through  $\bar{\theta}_n(\bar{\varrho})$ .

Finally, it is straightforward to obtain the derivatives of Eq. (5.6) in the main text for  $\bar{\varrho} = 0$ , i.e.,

$$\bar{\theta}'_n(0) = (n-1)\theta_n \left( \frac{1}{n-1} \right), \quad (5.57a)$$

$$\bar{\theta}''_n(0) = (n-1)(n-2) \left[ \theta_n \left( \frac{2}{n-1} \right) - 2\theta_n \left( \frac{1}{n-1} \right) \right]. \quad (5.57b)$$

Combining Eqs. (5.55), (5.56) and (5.57), we have self-consistent equations to solve for both the invasion threshold and the tricritical points.

### 5.9.1 Invasion threshold

We can identify the invasion threshold  $\beta_c$  for a fixed structure given by  $g_m$  and  $p_n$ , and for an arbitrary distribution  $\wp(\tau)$ . Combining Eqs. (5.55a), (5.56) and (5.57a), the constraint  $G'(0) = 1$  becomes

$$\left( \frac{\langle n(n-1)\theta_n \left( \frac{1}{n-1} \right) \rangle}{\langle n \rangle} \right) \left( \frac{\langle m^2 \rangle}{\langle m \rangle} \right) = \mu. \quad (5.58)$$

The left-hand side is intuitively interpreted as the average number of new infections caused by an infected node in a single time step near the absorbing state.

Let us assume that the conditions for the universal infection kernel are respected. Consequently, we can use

$$\theta_n(\varrho) \simeq D_\alpha \tau_c^\nu, \quad (5.59)$$

$$\propto [\beta f(n)\varrho]^\nu. \quad (5.60)$$

We keep the dependence on  $\beta$  and  $f(n)$  to see the impact of these parameters, whereas  $D_\alpha$  is just some constant that we cannot tune like the others. Consequently, we have  $\theta_n(1/n-1) \sim \beta^\nu [f(n)]^\nu (n-1)^{-\nu}$ , valid when  $\beta f(n)/(n-1) \ll 1$  and  $\mathcal{T}\beta f(n)/(n-1) \gg 1$ . We obtain

$$\beta_c \sim \left( \frac{\mu \langle n \rangle \langle m \rangle}{\langle n(n-1)^{1-\nu} [f(n)]^\nu \rangle \langle m^2 \rangle} \right)^{1/\nu}. \quad (5.61)$$

In the case  $f(n) = 1$ , we see that having  $\nu < 1$  induces a very small value for the invasion threshold.

### 5.9.2 Tricritical points

For a fixed structure given by  $g_m$  and  $p_n$ , the tricritical points  $(\beta_c, \nu_c)$  are solutions to both  $G'(0) = 1$  and  $G''(0) = 0$ . We substitute  $\bar{\theta}'(0) = \mu/\langle m \rangle$  in (5.55b) to obtain

$$G''(0) = \frac{1}{\mu} \left\{ \langle m^2 \rangle \bar{\theta}''(0) - \left( \langle m^2(m-1) \rangle + \frac{2\langle m^3 \rangle}{\mu} \right) \frac{\mu^2}{\langle m \rangle^2} \right\}. \quad (5.62)$$

Since the second term is strictly positive, it is necessary that  $\bar{\theta}''(0) > 0$  to have a tricritical point.

Using again the asymptotic expansion for  $\theta_n(\varrho)$  in combination with Eq. (5.56) and (5.57b), we get

$$\bar{\theta}''(0) \sim \frac{\beta^\nu}{\langle n \rangle} \langle n(n-1)^{1-\nu} (n-2) [f(n)]^\nu (2^\nu - 2) \rangle. \quad (5.63)$$

The tricritical point is evaluated at  $\beta = \beta_c$ , hence we can substitute the asymptotic expansion for the invasion threshold [Eq. (5.61)] to have an expression that only depends on  $\nu$ ,

$$\bar{\theta}''(0) \sim \mu \frac{\langle m \rangle \langle n(n-1)^{1-\nu} (n-2) [f(n)]^\nu (2^\nu - 2) \rangle}{\langle m^2 \rangle \langle n(n-1)^{1-\nu} [f(n)]^\nu \rangle}. \quad (5.64)$$

We get three insights from Eqs. (5.62) and (5.64).

1. It is necessary that  $\nu > 1$  to have a tricritical point, but not sufficient : it depends on the first three moments of  $g_m$ , and in a more complicated manner on the distribution  $p_n$ .
2. It is necessary to have  $p_n > 0$  for at least one value  $n > 2$ , i.e., environments of size  $n = 2$  cannot lead to a discontinuous phase transition.
3. A more heterogeneous  $g_m$  will typically require a larger  $\nu$  to reach a tricritical point. Indeed, if we keep  $\langle m \rangle$  fixed, but increase the value for the second and third moments (using a broader distribution for instance), the negative term on the right in Eq. (5.62) increases, while the positive term is invariant.

### 5.9.3 Participation time distribution with a fixed mean

In the main text, we fixed the minimal participation time to  $\tau_{\min} = 1$ , thereby fixing the scale of events relative to this minimal period, but we let  $\langle \tau \rangle$  vary with  $\alpha$ . Here we show the consequence of fixing  $\langle \tau \rangle = 1$ , and to let  $\tau_{\min}$  vary. For the sake of simplicity, we assume  $\mathcal{T} \rightarrow \infty$ . For a power-law distribution  $\wp(\tau) = C_\alpha \tau^{-\alpha-1}$ , this is achieved if we set  $\tau_{\min} = (\alpha - 1)/\alpha$ . Note that it is only possible for  $\alpha > 1$ , i.e., when the mean is well defined.

As shown in Fig. 5.7, keeping a fixed  $\langle \tau \rangle$  reveals another story for the bifurcation diagram. For a sufficiently large  $\beta$ , we now have that  $I^*$  grows with  $\alpha$ —a homogeneous participation time lead to



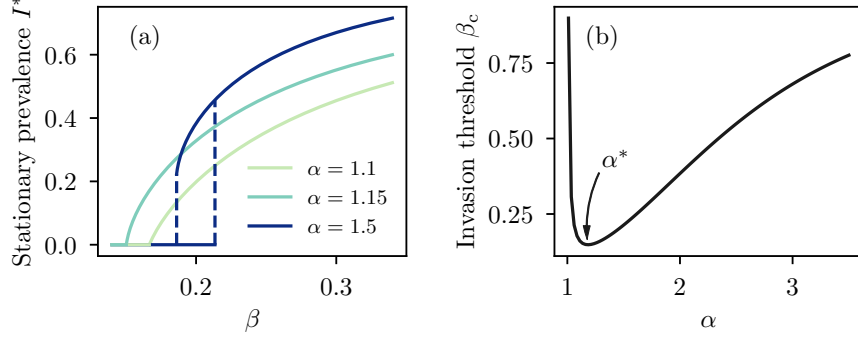


FIGURE 5.7 – Impact of a fixed mean participation time  $\langle \tau \rangle = 1$ . We use a power-law distribution  $\wp(\tau) \propto \tau^{-\alpha-1}$ ,  $\tau \in [(\alpha - 1)/\alpha, \infty)$ ,  $f(n) = 1$ , and  $\pi(\kappa; \lambda) \propto e^{-\kappa/\lambda}$  with  $K = 1$ . We use Poisson distributions for both  $g_m$  and  $p_n$ , with  $\langle m \rangle = 5$  and  $\langle n \rangle = 10$ , and set  $\mu = 0.05$ . (a) Phase diagram, highlighting the differences compared with a fixed  $\tau_{\min}$  as in Fig. 5.2 in the main text. For a sufficiently large  $\beta$ ,  $I^*$  now grows with  $\alpha$ . The dashed vertical line indicates the position of the persistence and invasion threshold for  $\alpha = 1.5$ . (b) The invasion threshold  $\beta_c$  is minimized for a value  $\alpha^* \approx 1.15$ .  $\beta_c$  goes to infinity in the limit  $\alpha \rightarrow 1$ , and converges to a certain value in the limit  $\alpha \rightarrow \infty$ .

bigger outbreaks—, which is the opposite of what we found for a fixed  $\tau_{\min}$ . Moreover, the invasion threshold is now a nonmonotonic function of  $\alpha$ . It is minimized for some value  $\alpha^* \approx 1.2$ , and the early spread of an epidemic would be maximized for this  $\alpha^*$ .

## 5.10 Appendix : Tracking dynamical correlations for quenched systems

In the main text, we supposed an annealed structure—nodes and groups are shuffled at each time step. This is why we can assume that the number of infected individuals in a group of size  $n$  is distributed according to a binomial distribution

$$H_{n,i} = \binom{n}{i} \bar{\varrho}^i (1 - \bar{\varrho})^{n-i}. \quad (5.65)$$

When calculating  $\bar{\theta}_n(\bar{\varrho})$  at Eq. (5.6), we restrict ourselves to groups with at least one susceptible node, hence we average  $\theta_n[i/(n-1)]$  over the distribution

$$\frac{(n-i)H_{n,i}}{\sum_i (n-i)H_{n,i}} = \frac{(n-i)}{(1-\bar{\varrho})n} H_{n,i} = \binom{n-1}{i} \bar{\varrho}^i (1-\bar{\varrho})^{n-i-1}, \quad (5.66)$$

which is again a binomial, but over the  $n-1$  remaining nodes.

In reality, people do engage with the same people in a recurrent manner, which can be represented by a *quenched structure*, i.e., an individual always participate to the same groups. In this case, we can track the evolution of  $H_{n,i}(t)$  using approximate master equations (AMEs) of the form

$$H_{n,i}(t+1) = \sum_{l=0}^{n-i} \sum_{j=0}^i B_{n+j-i-l,j}(\phi_{n,i+l-j}) B_{i+l-j,l}(\mu) H_{n,i+l-j}(t). \quad (5.67)$$

where

$$B_{u,v}(q) = \binom{u}{v} q^v (1-q)^{u-v}. \quad (5.68)$$

In Eq. (5.67),  $l$  counts the number of recoveries from  $t$  to  $t + 1$  and  $j$  counts the number of infections.  $\phi_{n,i} \equiv 1 - (1 - \theta_{n,i})(1 - \varphi)$  is the probability that a susceptible node belonging to a group  $(n,i)$  becomes infected. We have defined

$$\theta_{n,i} \equiv \theta_n \left( \frac{i}{n-1} \right), \quad (5.69)$$

which is the probability of getting infected by the group  $(n,i)$ , while  $\varphi$  is the average probability for a susceptible node to get infected through any *other* group to which it belongs. It corresponds to

$$\varphi(t) = \sum_m \frac{ms_m g_m}{\sum_{m'} m' s_{m'} g_{m'}} \left\{ 1 - [1 - \bar{\theta}(t)]^{m-1} \right\}. \quad (5.70)$$

The term in curly brace  $\{\dots\}$  is the probability for a susceptible node of membership  $m$  to get infected in any of the other  $m - 1$  groups to which it belongs. We average all this over a distribution  $\propto ms_m g_m$ , which is the probability for a random susceptible node in a group to be of membership  $m$ . Finally, the probability  $\bar{\theta}$  is now calculated as

$$\bar{\theta}(t) = \frac{\sum_{n,i} (n-i) \theta_{n,i} H_{n,i}(t) p_n}{\sum_{n,i} (n-i) H_{n,i}(t) p_n}. \quad (5.71)$$

Combining these equations with Eq. (5.4) in the main text, we have a closed system.

Note that this new set of equations still depends on the infection kernel  $\theta_{n,i}$ , hence our main result is directly transposed to this case where dynamical correlations are taken into account. In fact, the derivation of the infection kernel is agnostic to the underlying mathematical framework used to model the contagion.

It is also worth pointing out that we do not capture *exactly* all the local dynamical correlations with Eq. (5.67) : although we track  $H_{n,i}(t)$ , we assume that each susceptible node gets infected with independent probability  $\phi_{n,i}$ . However, for  $\mu \rightarrow 0$  and  $\phi_{n,i} \rightarrow 0$ , at most one node either recovers or becomes infected at the same time—in other words, we have asynchronous state transitions. In this limit (effectively taking the continuous time limit), we capture exactly the dynamical correlations within groups. Therefore, Eq. (5.67) can be seen as a good approximation for small  $\mu$  and  $\phi_{n,i}$ .

In Fig. 5.8, we show that the phenomenology of the contagion process using AMEs is very similar, almost indistinguishable from the phenomenology of the annealed system [see Fig. 5.2(b) and 5.2(c) in the main text]. Linear and sublinear kernel ( $\nu \leq 1$ ) still yield a standard exponential spread with a continuous phase transition, with a very small invasion threshold  $\beta_c$  for small  $\nu$ . Supralinear kernel ( $\nu > 1$ ) still leads to superexponential spread with a discontinuous phase transition.

It worth mentioning that even though the results presented here suggest that dynamical correlations within groups have a negligible impact on the contagion process, this does not hold for all scenarios (see Fig. 5.9). Moreover, in chapter 3 and 4, it was shown that these dynamical correlations are crucial for the onset of mesoscopic localization, which does significantly alter the phenomenology of the system. We expect dynamical correlations to have a larger impact for heterogeneous group size distribution  $p_n$  and more nonlinear contagions (larger  $\nu$ ).

## 5.11 Appendix : Comparison with simulations

To validate both the annealed equations and the AMEs, we performed simulations of the contagion process on random hypergraphs.

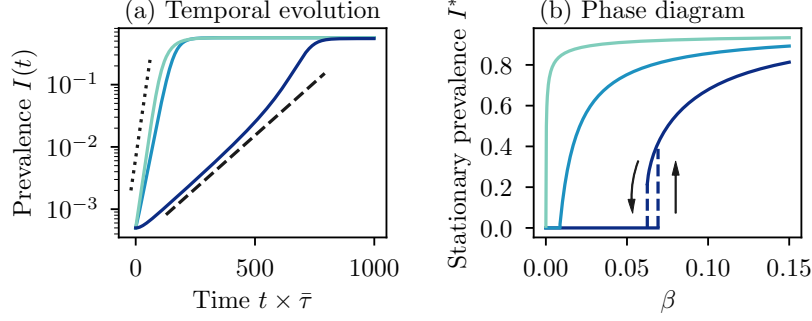


FIGURE 5.8 – Temporal evolution and phase diagram using the same conditions as Fig. 5.2 in the main text, but for a quenched system, making use of approximate master equations [Eqs. (5.67)-(5.71)].

### 5.11.1 Generation of random hypergraphs

We generate random hypergraphs with fixed membership and group size distributions  $g_m$  and  $p_n$ . First, we draw a membership sequence  $\mathbf{m} = [m_1, m_2, \dots, m_N]$  from  $g_m$ , where the number of nodes  $N$  is initially fixed. Second, we draw a group size sequence  $\mathbf{n} = [n_1, n_2, \dots, n_{N'}]$  from  $p_n$ , where the number of groups  $N'$  is not fixed. We want to have  $\sum_{j=1}^N m_j = \sum_{\ell=1}^{N'} n_\ell$ , therefore we add groups one at a time until  $\sum_{j=1}^N m_j \leq \sum_{\ell=1}^{N'} n_\ell$ . If  $\sum_{j=1}^N m_j < \sum_{\ell=1}^{N'} n_\ell$ , we remove groups randomly one at a time until  $\sum_{j=1}^N m_j \geq \sum_{\ell=1}^{N'} n_\ell$ . We alternate between these two processes until the equality is reached, which also determines  $N'$ .

Third, we assign the nodes to the groups, best interpreted as a bipartite graph realization. We create a stub list (list of half edges) for each sequence. For instance, for node  $j$ , we include  $m_j$  copies of the label  $j$  in the stub list of nodes. Similarly, we include  $n_\ell$  copies of group  $\ell$  in the stub list of groups. These two stub lists are of the same size by definition, hence we shuffle both and match the entries of the lists, thereby assigning nodes to groups at random.

However, this process does not avoid assigning a node  $j$  multiple times to a same group if  $m_j > 1$ . To remove these multiple assignments, we swap the concerned entries of the node stub list with another chosen at random, making sure it does not create another group with multiple assignments. In the bipartite representation, this move is equivalent to a double edge swap [55, 95]. We then perform a number of additional random double edge swaps ( $N$  in our simulations) to ensure the uniformity of the generation process.

### 5.11.2 Simulation of the contagion process

Let us denote a node  $j$  and the set of groups  $\ell$  to which it is assigned as  $\partial j$ . The parameters  $\alpha, T, \beta, f(n)$  and  $K$  are fixed. To simulate the contagion process, we start with an initial set of infected nodes  $\mathcal{I}(0)$  and a set of susceptible nodes  $\mathcal{S}(0)$  at time  $t = 0$ . At each time step  $t \rightarrow t+1$ , we perform the following operations.

1. We create a set of *new susceptible*  $\tilde{\mathcal{S}}(t+1)$ . For all  $j \in \mathcal{I}(t)$ , node  $j$  is included in set  $\tilde{\mathcal{S}}(t+1)$  with probability  $\mu$ .
2. We create a set of *new infected*  $\tilde{\mathcal{I}}(t+1)$ . For all  $j \in \mathcal{S}(t)$ , we iterate over groups  $\ell \in \partial j$ . For each group  $\ell$  we :
  - (i) draw a participation time  $\tau_\ell$ ;

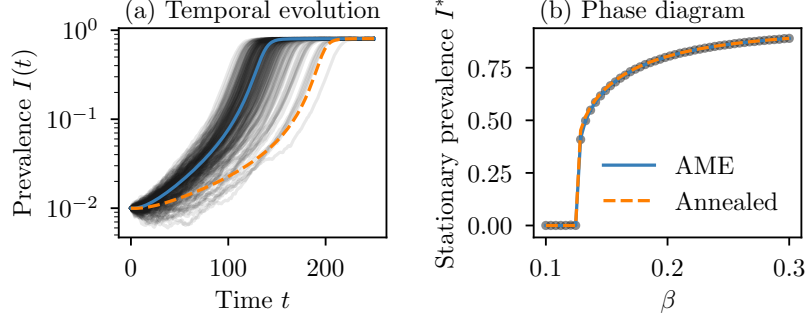


FIGURE 5.9 – Comparison with simulations for the temporal evolution and the phase diagram. We use an exponential dose distribution  $\pi(\kappa; \lambda) \propto e^{-\kappa/\lambda}$ , a power-law distribution of participation time  $\wp(\tau) \propto \tau^{-\alpha-1}$  with  $\alpha = 2$ , a clearing window  $\mathcal{T} \rightarrow \infty$ , and  $f(n) = 1$ . We set the recovery probability to  $\mu = 0.05$ . We use Poisson distributions for both  $g_m$  and  $p_n$ , with  $\langle m \rangle = 5$  and  $\langle n \rangle = 10$ . For the simulations, we use random hypergraphs with  $N = 10^4$  nodes. We use Eqs. (5.4)-(5.6) in the main text for the annealed solution and Eqs. (5.67)-(5.71) for the AMEs. (a) Temporal evolution. We performed 200 simulations each on a different hypergraph realization starting with a fraction  $I(0) = 10^{-2}$  of infected nodes. The semitransparent solid black lines show the prevalence  $I(t)$  for each realization. The AME solution is inside the region of high density of paths obtained from the simulations. (b) Phase diagram. For the simulations, we sampled the quasistationary state of the contagion process to estimate the stationary prevalence  $I^*$ . We performed measures on 10 hypergraph realizations. The semitransparent black circles correspond to the average prevalence and the vertical bars (too small to see) correspond to the standard deviation. Because we sample the quasistationary state, only the upper branch of the hysteresis loop is obtained. Both the annealed and AME solutions fit well with the simulations.

- (ii) given  $\tau_\ell$ , the group size  $m_\ell$ , and the fraction of the other nodes that are infected  $\varrho_\ell$ , we draw a dose  $\kappa_\ell$  from  $\pi(\kappa; \lambda_\ell)$ , where  $\lambda_\ell = \beta f(n) \tau_\ell \varrho_\ell$ ;
- 3. For all susceptible node  $j \in \mathcal{S}(t)$ , node  $j$  is added to the set  $\tilde{\mathcal{I}}(t+1)$  if  $\kappa_\ell > K$  for any of the group  $\ell \in \partial j$ .
- 4. We update the sets :  $\mathcal{I}(t+1) = \mathcal{I}(t) \cup \tilde{\mathcal{I}}(t+1) \setminus \tilde{\mathcal{S}}(t+1)$  and  $\mathcal{S}(t+1) = \mathcal{S}(t) \cup \tilde{\mathcal{S}}(t+1) \setminus \tilde{\mathcal{I}}(t+1)$ .

The prevalence at each time step is simply  $I(t) = |\mathcal{I}(t)|$ .

While the previous algorithm is sufficient to simulate the time evolution of the contagion process, estimating quantities in the stationary state can be problematic for finite size systems because the contagion process can get stuck in the absorbing state where all nodes are susceptible. We therefore use the *quasistationary-state method* [68] to avoid this problem. We keep a history of past states  $(\mathcal{I}, \mathcal{S})$ —50 states in our case—which we update after each decorrelation period  $t_d$ . To update the history, we replace a randomly chosen state in the history with the current state of the system. At any time during the simulation, if the absorbing state is reached, we replace the current state with one chosen randomly from the history.

To estimate quantities in the stationary state, we first let the system thermalize by performing  $10^3$  time steps. During this period, we use a short decorrelation period  $t_d = 1$  to make sure the whole history has been updated since the initial state. Then, we perform an additional  $10^3$  time steps during which we measure the system at regular intervals  $t_d = 10$ .

### 5.11.3 Comparison with analytical predictions

In Fig. 5.9(a), we present multiple realization of the stochastic process (semitransparent solid black lines) for the temporal evolution of the prevalence  $I(t)$ . The darker region corresponds to a high density of paths for  $I(t)$ . Interestingly, while both analytical approaches have a similar qualitative temporal evolution, AMEs (solid blue line) best capture the temporal evolution, with their prediction being in the region of high density of paths obtained from the simulations. Therefore, dynamical correlations do matter here. Moreover, since  $\alpha = 2$ , we have a supralinear infection kernel ( $\nu = 2$ ) leading to superexponential spread, a property validated by the simulations.

In Fig. 5.9(b), we compare the phase diagram obtained with our analytical approaches versus the simulations. In this case, both the annealed and the AMEs correctly predict the average prevalence in the stationary state—the prediction from AMEs are only slightly superior here. The simulations also confirm the discontinuous phase transition predicted by our analytical approaches.

## 5.12 Appendix : Code and data

We provide Python modules to solve for the infection kernel, the invasion threshold  $\beta_c$ , the time evolution of the system, both for the annealed and quenched systems (AMEs), and to perform simulations of the contagion process. The code is available here [263].

In the main text, Fig. 5.1(a) was created using data from Refs. [175, 272]. They provide collections of COVID-19 superspreading events that were either reported in news outlet, on government websites, or in research papers.

We kept those where the attack rate, the duration of the event, and the size of the event (number of people attending) were either reported or could be calculated. The curated data is available here [264].

## Chapitre 6

# Sur l'importance des groupes influents pour les contagions non linéaires

Article original :

**Influential groups for seeding and sustaining nonlinear contagion in heterogeneous hypergraphs**

**Guillaume St-Onge**, Iacopo Iacopini, Vito Latora, Alain Barrat, Giovanni Petri, Antoine Allard et Laurent Hébert-Dufresne

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

Référence : Commun. Phys. **5**, 25 (2022) [262]

© 2021 Nature (§6.3-6.11)<sup>1</sup>

---

1. Ces sections contiennent le contenu original de l'article. Celui-ci n'a été modifié que pour se conformer au format exigé par la Faculté des études supérieures et postdoctorales de l'Université Laval

## 6.1 Avant-propos

Dans ce chapitre, nous caractérisons en profondeur les contagions complexes sur réseaux d'ordre supérieur grâce aux équations maîtresses approximatives. Nous nous intéressons plus particulièrement aux taux d'infection non linéaires, dénotés par  $\beta(n, i)$ , où  $n$  est la taille d'un groupe et  $i$  est le nombre de noeuds infectés dans le groupe.

Afin d'affiner notre exploration, nous utilisons les résultats du chapitre précédent pour motiver une forme particulière de taux d'infection. Au chapitre 5, nous avons démontré l'existence d'une forme universelle pour  $\theta_n(\varrho)$ , la probabilité qu'un noeud se fasse infecter via un groupe de taille  $n$ , sachant qu'une fraction  $\varrho = i/(n-1)$  des autres noeuds du groupe sont contagieux. La forme universelle de  $\theta_n(\varrho)$  correspond au développement asymptotique

$$\theta_n(\varrho) \sim \tau_c^{-\nu} \propto [f(n)\varrho]^\nu,$$

où  $f(n)$  permet d'ajuster la fréquence des contacts à l'intérieur d'un groupe en fonction de la taille  $n$ . Si la probabilité d'infection  $\theta_n \ll 1$ , alors il est possible d'utiliser la limite du temps continu  $t \in \mathbb{R}^+$ , et donc  $\theta_n$  peut s'interpréter comme un taux d'infection  $\theta_n(\varrho) \mapsto \beta(n, i)$ .

Lorsque la fréquence des contacts est proportionnelle au nombre de voisins dans le groupe,  $f(n) \propto n-1$ , alors  $\beta(n, i)$  est indépendant de  $n$  :

$$\beta(n, i) \simeq \lambda i^\nu.$$

C'est le taux d'infection non linéaire que nous utilisons dans ce chapitre. Pour une fonction  $f(n)$  arbitraire, il suffirait de considérer  $\beta(n, i) \simeq \Lambda(n)i^\nu$ .

Notre étude est reliée de près au modèle de la *contagion simpliciale* [142] présenté au chapitre 2. Ce modèle fut d'ailleurs repris dans plusieurs articles récents [20, 40, 56, 59, 88, 147, 170]. Toutefois, l'approche par équations maîtresses approximatives nous permet de pousser plus loin l'analyse de ce modèle sur les hypergraphes ayant une structure hétérogène. Notamment, nous mettons en évidence le rôle primordial joué par certains groupes dans le maintien d'un état endémique stable ou dans la vitesse de la progression d'une épidémie dans une population.

Symbole	Description
$n$	Taille de groupe : nombre de noeuds dans un groupe
$m$	Appartenance : nombre de groupes auxquels un noeud appartient
$n_{\max}$	Taille de groupe maximale
$m_{\max}$	Appartenance maximale
$p_n$	Distribution des tailles de groupe
$g_m$	Distribution des appartenances
$\gamma_n$	Exposant de la loi de puissance $p_n \propto n^{-\gamma_n}$
$\gamma_m$	Exposant de la loi de puissance $g_m \propto m^{-\gamma_m}$

$i$	Nombre de noeuds infectés dans un groupe
$f_{n,i}(t)$	Fraction des groupes de taille $n$ possédant $i$ noeuds infectés au temps $t$
$s_m(t)$	Fraction des noeuds d'appartenance $m$ qui sont susceptibles au temps $t$
$I(t)$	Prévalence : fraction des noeuds qui sont infectés au temps $t$
$I_n(t)$	Fraction moyenne des noeuds qui sont infectés dans un groupe de taille $n$ au temps $t$
$\beta(n, i)$	Taux d'infection général pour un noeud susceptible exposé à un groupe de taille $n$ avec $i$ noeuds infectés
$\lambda$	Coefficient de proportionnalité pour le taux d'infection particulier $\beta(n, i) = \lambda i^\nu$
$\nu$	Exposant pour le taux d'infection particulier $\beta(n, i) = \lambda i^\nu$
$r(t)$	Taux d'infection moyen par groupe auquel un noeud susceptible est exposé au temps $t$
$\rho(t)$	Taux d'infection moyen pour tous les <i>autres</i> groupes auquel un noeud susceptible est exposé au temps $t$ , si on choisit le noeud aléatoirement dans un groupe
$h_{n,i}$	Dérivée de $f_{n,i}^*$ par rapport à $\rho^*$ lorsque $\rho^* \rightarrow 0^+$
$\lambda_c$	Seuil d'invasion : $I^* = 0$ est une solution instable pour $\lambda > \lambda_c$
$\lambda_p$	Seuil de persistance : $I^* = 0$ la seule solution stable pour $\lambda < \lambda_p$
$\nu_c$	Seuil de bistabilité : la transition de phase sur $\lambda$ est discontinue avec un régime bistable pour $\nu > \nu_c$
$\mathcal{S}$	Ensemble des conditions initiales pour les noeuds : $\mathcal{S} = \{s_m(0)\}_{m=1}^{m_{\max}}$
$\mathcal{F}$	Ensemble des conditions initiales pour les groupes : $\mathcal{F} = \{f_{n,i}(0)   0 \leq i \leq n\}_{n=2}^{n_{\max}}$
$\Phi[\mathcal{S}, \mathcal{F}]$	Fonction objectif pour maximiser la croissance initiale de la prévalence $\dot{I}(0)$

TABLEAU 6.1 – Glossaire des symboles importants utilisés au chapitre 6.

## 6.2 Résumé

Plusieurs aspects des contagions de nature biologique et sociale sont le résultat d'interactions entre plusieurs individus, comme les évènements de superdissémination ou la pression sociale exercée pour qu'un individu adopte un comportement. Il est alors plus naturel de représenter les interactions sous la forme d'un hypergraphe. Dans cet article, nous développons un nouveau cadre théorique basé sur les équations maîtresses approximatives pour étudier les processus de contagion sur des hypergraphes aléatoires. Notre approche permet d'étudier les structures hétérogènes, autant au niveau de la taille des groupes (cardinalité des hyperliens) que de l'appartenance des noeuds aux groupes (hyperdegré). Puisque nous caractérisons de manière quasi exacte la dynamique interne des groupes, nos prédictions sont fidèles aux simulations, et ce, tout en procurant une description mathématique suffisamment simple qui nous permet de démontrer des résultats analytiques. En utilisant un modèle pour lequel les



interactions d'ordre supérieur mènent à un taux d'infection non linéaire, nos deux résultats principaux montrent que les groupes de grande taille sont influents, au sens où ils sont responsables de la croissance initiale de la prévalence et permettent à eux seuls de maintenir un état endémique. Dans un premier temps, nous présentons une analyse détaillée de la transition de phase du modèle, qui peut être continue ou discontinue avec un régime de bistabilité. Nous obtenons des expressions analytiques pour les points critiques et tricritiques, ce qui nous permet d'établir qu'une grande valeur pour le troisième moment de la distribution des appartenances empêche l'apparition d'une transition de phase discontinue. De plus, la non-linéarité de la contagion et l'hétérogénéité de la taille des groupes favorisent l'émergence de la localisation mésoscopique, auquel cas les groupes de grande taille maintiennent l'état endémique, et par le fait même nuisent à l'apparition d'une phase bistable. Dans un second temps, nous introduisons un problème d'optimisation pour les conditions initiales dans le but de favoriser une invasion rapide du système. Nous comparons deux stratégies : choisir les conditions initiales en fonction des propriétés des noeuds ou en fonction des propriétés des groupes. Quand le processus de contagion est suffisamment non linéaire, nous montrons qu'il est plus important de s'attarder aux propriétés des groupes pour envahir rapidement le système.

### 6.3 Abstract

Several biological and social contagion phenomena, such as superspreading events or social reinforcement, are the results of multibody interactions, for which hypergraphs offer a natural mathematical description. In this chapter, we develop a novel mathematical framework based on approximate master equations to study contagions on random hypergraphs with a heterogeneous structure, both in terms of group size (hyperedge cardinality) and of membership of nodes to groups (hyperdegree). The characterization of the inner dynamics of groups provides an accurate description of the contagion process, without losing the analytical tractability. Using a contagion model where multibody interactions are mapped onto a nonlinear infection rate, our two main results show how large groups are influential, in the sense that they drive both the early spread of a contagion and its endemic state (i.e., its stationary state). First, we provide a detailed characterization of the phase transition, which can be continuous or discontinuous with a bistable regime, and derive analytical expressions for the critical and tricritical points. We find that large values of the third moment of the membership distribution suppress the emergence of a discontinuous phase transition. Furthermore, the combination of heterogeneous group sizes and nonlinear contagion facilitates the onset of a mesoscopic localization phase, where contagion is sustained only by the largest groups, thereby inhibiting bistability as well. Second, we formulate the problem of optimal seeding for hypergraph contagion, and we compare two strategies: tuning the allocation of seeds according to either individual node or group properties. We find that, when the contagion is sufficiently nonlinear, groups are more effective seeds of contagion than individual nodes.

### 6.4 Introduction

Mathematical models of contagion processes enhance our understanding of spreading dynamics and our predictive capabilities [12, 157]. To account for the interconnected nature of real-world systems, the last two decades of network science and computational epidemiology research have focused on modeling frameworks of increasing complexity [19, 163, 231]. From the spreading of infectious diseases to rumors and innovations [66, 208, 244], a crucial aspect remains the interplay between the contact

patterns enabling transmission and the dynamics that unfolds on top. As a representation for these contacts, graphs have been widely exploited to better represent real-world patterns with increasing levels of accuracy [19, 220, 231]. When adequate, static interactions have been replaced with their weighted, time-varying, and multilayered counterparts, respectively accounting for the differences in their strength and duration, for their temporally resolved nature, and for the existence of different types of interactions [18, 35, 134].

While graphs remain a reference for the representation of complex systems, they come with a fundamental limitation : they can only encode pairwise interactions (represented by edges). Groups are instead the foundational units of many biological, ecological and social systems, whose processes can involve multibody interactions between any number of elements. To overcome this limitation, higher-order network representations [277] are becoming more and more popular [22, 169, 247] to describe the structure of interacting systems [49, 238], their growth [31, 65, 119, 299] and dynamics in groups [125, 137, 226]. Simplicial complexes [117] and hypergraphs [27] can encode relationships and interactions between any number of elements. They have been used to investigate the implications of higher-order interactions for landmark dynamical processes like synchronization [32, 199, 257], diffusion [43, 250, 276], and other social dynamics [9, 141, 216]. Ref. [22] provides a comprehensive review of early efforts in this direction.

Recently, a model of simplicial contagion has been proposed [142]. This is a standard Susceptible-Infected-Susceptible (SIS) compartmental model in which a susceptible individual can become infected through different transmission channels, beyond infectious edges. In models of *simple contagion*, the transition from susceptible to infected happens independently with each exposure to an infectious edge [231]. In models of *complex contagion* instead [52], the transition requires multiple infectious edges or is reinforced by multiple exposures, thus accounting for the empirically observed mechanisms of social influence [50, 130, 151, 176, 207]. In simplicial contagions, or more generally *hypergraph contagions*, a susceptible individual can become infected because of a multibody process, i.e., through an exposure to an infectious group [142]. In this way, the node is simultaneously exposed to the state of the entire group, whose effect can be interpreted as a mechanism of social reinforcement [287]. In addition, the study of higher-order contagion models has applications as well in biological sciences : it has indeed recently been shown that the combination of multibody interactions, heterogeneous temporal activity, and the concept of a minimal infective dose lead to nonlinear infection kernels in a model of biological contagions [265].

The analytical approaches derived so far have confirmed the rich phenomenology emerging from the contagion dynamics, characterized by discontinuous phase transitions, bistability and critical mass effects [20, 59, 90, 142, 147, 170]. Most approaches follow a heterogeneous mean-field (HMF) framework in which nodes are divided into hyperdegree classes. These descriptions are analytically tractable, but do not consider the details of the structure and ignore the *dynamical correlations* within groups, which are especially important for hypergraph contagions since multibody interactions naturally reinforce these correlations.

Other approaches like quenched mean-field theory (QMF) [88] and microscopic Markov-chains (MMCA) [56, 196] can explicitly take the entire contact patterns into account. Along this line, the microscopic epidemic clique equations (MECLE) capture dynamical correlations as well, thereby highlighting the important impact these correlations have on critical points [40]. The analytical tractability of these

approaches is, however, sacrificed in favor of a more precise description of the structure. To fully understand the consequences of multibody interactions in contagions on higher-order networks, we thus need a framework that is both analytically tractable and captures dynamical correlations.

In particular, we are interested in better understanding the notion of influence in hypergraph contagions. In classic contagion models on random networks, individual hubs are influential in the sense that they are the best seeds of contagions, but they are also the most apt at sustaining seeded contagions [230]. However, in hypergraph contagions we must consider the influence of both individuals and groups, because dynamical correlations can allow groups to be more influential than sets of uncorrelated hubs. Regimes of bistability and hysteresis also imply a potential decoupling between the ability of nodes to seed a contagion and their ability to sustain it. We thus wish to determine (i) which set of groups can best maintain the stationary state of a hypergraph contagion, and when this becomes a dominant effect (ii) which set of groups and their configuration offer optimal initial conditions for a contagion, as compared to the classic notion of influential spreaders as individual hubs, and (iii) whether or not these two notions of group influence are aligned.

We use approximate master equations (AMEs) [105, 125, 189, 226, 266, 267] to study hypergraph contagions, capturing exactly the inner dynamics of groups. We apply this framework to a contagion model where the infection rate is a nonlinear function of the number of infected nodes in groups, and show that *influential groups* can drive both the stationary state of the contagion and its behavior in the transient state.

The chapter is structured as follows. In Sec. 6.5, we map the simplicial contagion model onto a hypergraph contagion process and introduce the mathematical formalism of group-based AMEs [125, 266, 267]. To illustrate the accuracy of our framework and assess its limitations, we compare the predictions of our AMEs with simulations on real-world hypergraphs and their randomized counterparts.

In Sec. 6.6, we characterize the importance of influential groups on the phase transition. We derive analytical expressions for the critical and tricritical points, and study the impact of a heterogeneous structure. We show that large values for the third moment of the *membership* (hyperdegree) distribution suppress the emergence of a discontinuous phase transition, a result consistent with HMF theories [147, 170]. Our formalism allows us to go further and capture the emergence of a *mesoscopic localization* phase, where infected nodes are concentrated in the largest groups (see chapter 3 and 4). We show that localization effects can drive the onset of an endemic phase and, incidentally, inhibit bistability.

Finally, in Sec. 6.7, we showcase the importance of influential groups on the transient state by considering a problem of influence maximization. We define and solve the optimization problem based on two strategies, allocating seeds according to either node individual properties or according to group properties. When the contagion is sufficiently nonlinear, we find that the latter is more effective. This suggests again that contagion processes driven by *groups* lead to a fundamentally different phenomenology than pairwise contagions, such that investigations and modeling of social contagions should carefully account for higher-order interactions.

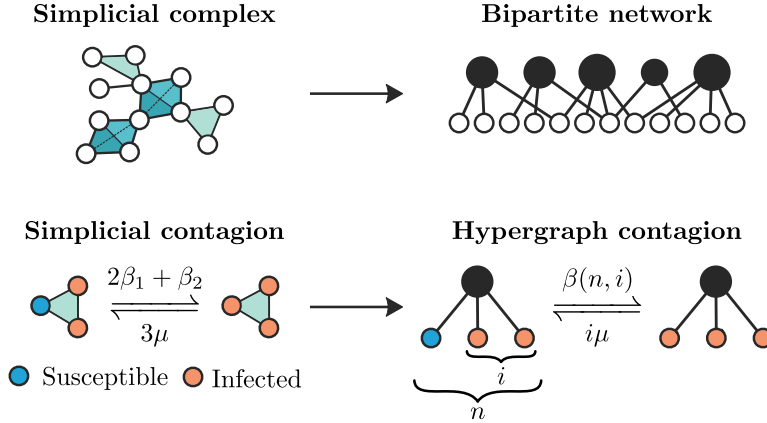


FIGURE 6.1 – Mapping of the simplicial contagion model to a hypergraph contagion. We use a bipartite representation, where nodes (white circles) belong to *groups* (black circles). A facet of dimension  $n - 1$  is mapped onto a group of size  $n$ . In the simplicial contagion model, contributions from higher-order interactions are taken into account by additional transmission rates (e.g.,  $\beta_2$  for the 2-simplex) when all but one node of the simplex are infected [142]. With our hypergraph representation, infections within a group are simply modeled by a general infection function  $\beta(n, i)$  that depends on both the size  $n$  of the group and the number  $i$  of infected nodes in the group (with  $i \leq n$ ).

## 6.5 Hypergraph contagion model

In the original version of the simplicial contagion model [142], the spreading process takes place on a simplicial complex [see Fig. 6.1] and obeys the following rule. If all nodes in a  $d$ -simplex are infected except a susceptible one, this remaining node gets infected at a rate  $\beta_d$ , and also receives contributions from the lower-dimensional simplices included in the  $d$ -simplex with rates respectively equal to  $\beta_{d-1}, \dots, \beta_1$ . For instance, in Fig. 6.1, two of the three nodes composing a 2-simplex are infected, hence the susceptible node gets infected at rate  $2\beta_1 + \beta_2$ , considering the contributions both from the two edges and from the “triangle”.

A simplicial complex is a specific type of hypergraph, and thus we can always map the former on the latter—note that the reverse direction is not always possible. To do so, each facet—a simplex that is not a face of any larger simplex—is represented by a single hyperedge (group). In this chapter, we relax the requirement of the simplicial complex in favor of a more general hypergraph structure. We make use of the bipartite representation of hypergraphs [22, 277], in which the two sets of nodes of the bipartite graph correspond respectively to the sets of nodes and groups of the original hypergraph, as illustrated in Fig. 6.1.

The *size*  $n$  of a group corresponds to the number of nodes belonging to this group, and it is therefore equivalent to the *hyperedge cardinality*. Note that a  $d$ -simplex consists of  $d + 1$  nodes and is therefore mapped onto a group of size  $n = d + 1$ . Similarly, the *membership*  $m$  of a node, the *node hyperdegree*, corresponds to the number of groups to which it belongs, regardless of their size.

The hypergraph contagion model is defined as follows : for a group of size  $n$ , where  $i \leq n$  members are infected, each of the  $n - i$  susceptible nodes gets infected at rate  $\beta(n, i)$ . For susceptible nodes that belong to multiple groups, their total transition rate to the infected state is simply the sum of the infection rates associated with each group to which they belong—in other words, the infection

processes are independent. All infected nodes transition back to the susceptible state at the same constant recovery rate  $\mu$ .

Notice that the hypergraph contagion model allows to represent any type of simplicial contagion. For instance, in the simplicial contagion model case of Fig. 6.1, we would use  $\beta(3, 2) = 2\beta_1 + \beta_2$ . In fact, the description offered by the infection rate function  $\beta(n, i)$  yields a variety of models more general than the original simplicial contagion—in which a function  $\beta(i)$  would be sufficient to encode contributions from facets of any dimension.

In all our case studies, we will use an infection rate function of the form

$$\beta(n, i) = \lambda i^\nu . \tag{6.1}$$

However, many results we derive hold true for a general infection rate function  $\beta(n, i)$ . The parameter  $\nu$  controls the *nonlinearity* of the contagion. A linear contagion is recovered by setting  $\nu = 1$ , which is equivalent to a standard SIS model on networks, where each group is a clique [125]. We intentionally chose the infection rate function independent of  $n$  to focus on the impact of a nonlinear dependence on  $i$ ; it would be straightforward to generalize the results by considering  $\beta(n, i) \mapsto \Lambda(n)i^\nu$ .

The infection rate function in Eq. (6.1) is the simplest nonlinear generalization of standard epidemiological models, where  $\beta(n, i) \propto i$ . Moreover, we can motivate the choice of exponents  $\nu \neq 1$  in the context of social contagions, by comparing our approach to the original formulation of the simplicial contagion model. A value of  $\beta_2 > 0$  in Fig. 6.1 represents *social reinforcement* [142], and to correctly map the infection rate for a triangle, we need to use an exponent  $\nu > 1$  in our model. Similarly, a value  $\beta_2 < 0$  represents *social inhibition*, and this case can be obtained with an exponent  $\nu < 1$ .

Another motivation for the infection rate function at Eq. (6.1) is a recent study that shows this general form emerges in the occurrence of heterogeneous temporal patterns [265]. More specifically, if you consider that the participation time of nodes—representing individuals—to higher-order interactions is distributed according to a power law, and that individuals become infected according to a threshold mechanism based on the *dose* received in the interaction, then the probability for a node to get infected in a group is  $\propto i^\nu$ , where  $\nu$  is related to the temporal heterogeneity. In the continuous time limit, one recovers the infection rate function defined at Eq. (6.1).

In Ref. [265], the infection mechanism is motivated in the context of biological contagions, where the infective dose received could represent viral particles for instance, and the threshold would correspond to the minimal infective dose to develop a disease. While such type of complex contagions are rarely used in the context of biological contagions, they could help explain certain observed phenomena, such as super-exponential spread for certain diseases [249]. Moreover, threshold models are very common in social contagions [50, 52, 75, 112, 288], thus Eq. (6.1) could be interpreted as an effective mechanism of social spread accounting for heterogeneous temporal patterns.

### 6.5.1 Group-based Approximate Master Equations

To describe hypergraph contagions, we make use of group-based Approximate Master Equations [125, 226, 266, 267]. This means that we do not rely on specific hypergraph realizations. Instead, we assume that the structure is drawn from a random hypergraph ensemble described by the distributions  $p_n$ , for the size  $n$  of a group, and  $g_m$ , for the membership  $m$  of a node. Each of the  $m$  membership stubs of

a node is assigned uniformly at random to a group available spot. Therefore, the membership  $m$  of a node and the sizes of the groups to which it belongs are uncorrelated.

To track the evolution of a contagion process on this ensemble of hypergraphs, we define two sets of quantities :  $s_m(t)$ , representing the fraction of nodes with membership  $m$  that are susceptible at time  $t$  and  $f_{n,i}(t)$ , the fraction of groups of size  $n$  having  $i$  infected members at time  $t$ . The last quantity can also be interpreted as a conditional probability (of observing  $i$  infected nodes in a group of size  $n$ ) satisfying the normalization condition  $\sum_i f_{n,i} = 1$ .

We further define two mean-field quantities. First, let us take a random susceptible node. The mean-field infection rate resulting from a random group to which it belongs is defined as

$$r(t) = \frac{\sum_{n,i} \beta(n,i)(n-i)f_{n,i}p_n}{\sum_{n,i} (n-i)f_{n,i}p_n} . \quad (6.2)$$

Indeed, the joint distribution for the size  $n$  and the number of infected nodes  $i$  in this group is proportional to  $(n-i)f_{n,i}p_n$ , and we just average  $\beta(n,i)$  over this distribution.

Second, let us randomly choose a susceptible node inside a group. The mean-field infection rate caused by all the *external* groups to which the susceptible node belongs (excluding the one from which we picked the node) can be written as

$$\rho(t) = r(t) \frac{\sum_m m(m-1)s_m g_m}{\sum_m m s_m g_m} . \quad (6.3)$$

To obtain  $\rho(t)$ , we assume that infections coming from different groups are independent processes. We multiply  $r(t)$  with the mean *excess* membership of a susceptible node, i.e., if we pick a susceptible node in a group, it is the expected number of *other* groups to which it belongs. Since the membership distribution of a susceptible node picked in a group is proportional to  $m s_m(t) g_m$ , we simply average  $m-1$ , its excess membership, over this distribution.

Using the definitions in Eqs. (6.2) and (6.3) we can write the following system of AMEs

$$\frac{ds_m}{dt} = \mu(1 - s_m) - m r s_m , \quad (6.4a)$$

$$\begin{aligned} \frac{df_{n,i}}{dt} = & \mu(i+1)f_{n,i+1} - \mu i f_{n,i} \\ & - (n-i) [\beta(n,i) + \rho] f_{n,i} \\ & + (n-i+1) [\beta(n,i-1) + \rho] f_{n,i-1} . \end{aligned} \quad (6.4b)$$

This system is composed of  $O(n_{\max}^2 + m_{\max})$  equations. It is approximate in the sense that the evolution of the fractions of infected nodes of membership  $m$ ,  $s_m$ , are treated in a mean-field fashion (still considering dynamic correlations between pairs node-group), while the evolution equations of the  $f_{n,i}$  are treated as master equations. In the right-hand side of Eq. (6.4b), the first two terms are due to the recovery process, and the last two to the infection. The infection rate due to infected nodes *inside* the group is exact, while the infection rates associated to *external* groups (the terms making use of  $\rho$ ) are treated in an approximate way. Without loss of generality (up to a change of time scale) we set  $\mu \equiv 1$  for the remainder of the chapter.

From Eq. (6.4), we can calculate the global prevalence

$$I(t) = \sum_m [1 - s_m(t)] g_m , \quad (6.5)$$

and the group prevalence

$$I_n(t) = \sum_i \frac{i}{n} f_{n,i}(t), \quad (6.6)$$

which correspond to the average fraction of infected nodes in the whole system and within groups of size  $n$  respectively.

In the stationary state, we obtain the following self-consistent relations :

$$s_m^* = \frac{1}{1 + mr^*}, \quad (6.7a)$$

$$(i+1)f_{n,i+1}^* = \{i + (n-i)[\beta(n,i) + \rho^*]\} f_{n,i}^* - (n-i+1)[\beta(n,i-1) + \rho^*] f_{n,i-1}^*. \quad (6.7b)$$

The latter equation can be simplified by noting that  $f_{n,i}^*$  must respect *detailed balance* in the stationary state, i.e.,

$$(i+1)f_{n,i+1}^* = (n-i)[\beta(n,i) + \rho^*] f_{n,i}^*,$$

expressing that the probability to decrease the number of infectious nodes in a group of size  $n$  from  $i+1$  to  $i$  by a recovery process is equal to the probability of the opposite change (from  $i$  to  $i+1$  infectious nodes) obtained through a contagion event. We thus finally obtain

$$f_{n,i}^* = f_{n,0}^* \frac{n!}{(n-i)!i!} \prod_{j=0}^{i-1} [\beta(n,j) + \rho^*] \quad \forall i \in \{1, \dots, n\}, \quad (6.8)$$

where  $f_{n,0} = 1 - \sum_{i>0} f_{n,i}$ .

### 6.5.2 Comparison with simulations

We provide a comparison of our approach with Monte Carlo simulations (see Methods Sec. 6.9.1). To do this, we consider empirical higher-order structures constructed from real-world data, and their randomized counterparts. Details on the data collection and aggregation are provided in Methods Sec. 6.9.2.

The motivation for this *a priori* validation is threefold : first, it allows us to illustrate the validity and accuracy of our analytical framework when our assumptions are met—namely when the structure is drawn from an ensemble of uncorrelated hypergraphs with fixed  $g_m$  and  $p_n$ . Second, because of the excellent agreement with simulations for random hypergraphs, we omit further comparison with Monte Carlo simulations for the many results we present in the following sections. Finally, it showcases the possible sources of discrepancies—and how our results could vary—when considering real datasets.

In Fig. 6.2, we show the phase diagram, i.e., the fraction  $I^*$  of infectious nodes in the stationary state, of contagion dynamics on hypergraphs that encode higher-order social (face-to-face) interactions between individuals from a primary school in Lyon (see Methods Sec. 6.9.2 for details). Both the membership and group size distributions are homogeneous for this dataset. We considered linear contagion ( $\nu = 1$ ), equivalent to the standard SIS model, and superlinear contagion ( $\nu = 4$ ). In both cases, our analytical formalism (continuous lines) agrees quite well with the Monte Carlo simulations (symbols) on the original (empirical) hypergraph [Figs. 6.2(a) and 6.2(d)]. The main source of errors for the observed

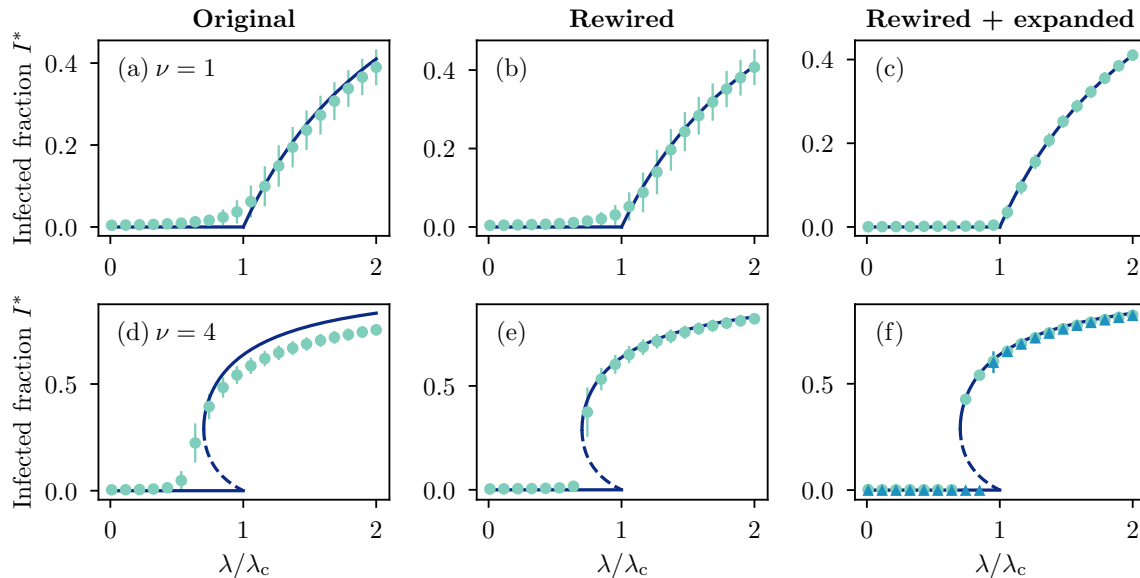


FIGURE 6.2 – Fraction  $I^*$  of infectious nodes in the stationary state of a contagion model on a hypergraph constructed from high-resolution face-to-face contact data from a primary school in Lyon (see Methods Sec. 6.9.2). The hypergraph contains 242 nodes and 1188 groups. Both the membership and the group size distributions are homogeneous, with  $\langle m \rangle \approx 11.79$ ,  $\langle n \rangle \approx 2.40$ ,  $m_{\max} = 32$  and  $n_{\max} = 5$ . We compare the results of Monte Carlo simulations (symbols; see Methods Sec. 6.9.1) with the predictions of our approach (solid and dashed lines for stable and unstable solutions respectively). We rescale  $\lambda$  with the invasion threshold  $\lambda_c$ , which is computed using Eq. (6.15) in Sec. 6.6. The symbols represent the average infected fraction measured over long runs and averaged over randomized hypergraphs when this applies. The error bars (sometimes too small to be seen) correspond to one standard deviation. The green circles are obtained with the quasistationary-state method, starting with a large fraction of infected nodes ( $I = 0.8$ ) to sample the upper branch of the hysteresis loop when the phase transition is discontinuous. (a),(d) We use the original hypergraph. (b),(e) We use 10 uniformly randomized versions of the original hypergraph. (c),(f) We use 10 uniformly randomized versions of the original hypergraph with its size expanded by a factor 10 (see Methods Sec. 6.9.3). (f) The blue triangles are obtained by ordinary simulations, starting with a small fraction of infected nodes ( $I = 0.02$ ) to sample the lower branch of the hysteresis loop. We only do it for the expanded hypergraphs because finite-size effects makes unreliable the estimation of the lower branch with data of small size.

discrepancy can be attributed to *structural* correlations, which do not appear to affect the threshold values, but reduce the stationary prevalence. Indeed, in Figs. 6.2(b) and 6.2(e), the agreement improves by randomizing the hypergraph while preserving the memberships and the group sizes. The remaining discrepancies are due to the fact that simulations are affected by finite-size effects, while our formalism assumes an infinite size system : the agreement becomes almost perfect in Figs. 6.2(c) and 6.2(f) by additionally increasing the size of the hypergraph by a factor 10.

Let us remark that the error is more important for  $\nu = 4$  on the original hypergraph [Fig. 6.2(d)], which suggests that structural correlations have a greater effect on nonlinear contagions. In the Supplementary Information, we show how to generalize our framework to account for structural correlations.

We also considered a completely different dataset, which represents co-authorship relations in computer science publications, obtained from major journals and proceedings in the field. The resulting



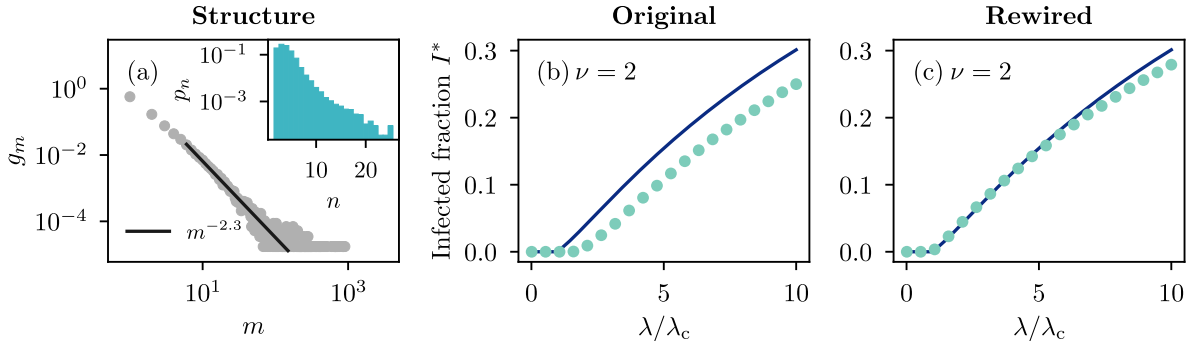


FIGURE 6.3 – Fraction  $I^*$  of infectious nodes in the stationary state of a contagion model on a hypergraph constructed from coauthorship data in computer science (see Methods Sec. 6.9.2). The hypergraph has been obtained by using a breadth-first search to extract a sub-hypergraph of 57 501 nodes and 55 204 groups; the original data contained more than  $10^6$  nodes and groups. (a) The membership distribution is heterogeneous, of the form  $g_m \sim m^{-\gamma_m}$  for large  $m$ , while the group size distribution is more homogeneous, with  $\langle m \rangle \approx 3.75$ ,  $\langle n \rangle \approx 3.90$ ,  $m_{\max} = 903$  and  $n_{\max} = 25$ . (b),(c) We compare the results of Monte Carlo simulations (circle markers; see Methods Sec. 6.9.1) with the predictions of our approach (solid lines). We rescale  $\lambda$  with the invasion threshold  $\lambda_c$ , which is computed using Eq. (6.15) in Sec. 6.6. The infected fraction  $I^*$  has been obtained as averages over time with long runs (and averaging over randomized hypergraphs when this applies). The error bars (too small to be seen) correspond to one standard deviation. We used ordinary simulations, starting with  $I = 0.8$ . (b) We use the original hypergraph. (c) We use 10 uniformly randomized versions of the original hypergraph.

hypergraph is considerably larger than the previous one, and it also presents a very different structure (see Methods Sec. 6.9.2). The results are shown in Fig. 6.3, where we plot the same phase diagram curves as in Fig. 6.2, using a superlinear contagion ( $\nu = 2$ ). In this case, however, the membership distribution is heterogeneous [Fig. 6.3(a)], of the form  $g_m \sim m^{-\gamma_m}$  for large  $m$  with  $\gamma_m \approx 2.3$ , while the group size distribution is more homogeneous [inset of Fig. 6.3(a)], but still extends to rather large values with  $n_{\max} = 25$ . By comparing the results for the original hypergraph [Fig. 6.3(b)] against those for a randomized ensemble [Fig. 6.3(c)], we see that structural correlations account for the major part of the discrepancies between simulations and theory, affecting both the invasion threshold and the stationary prevalence. However, Fig. 6.3(c) shows that structural correlations are not the only source of errors at high prevalence.

Let us recall that our formalism correctly captures the *dynamical* correlations within a group through a master equation description [Eq. (6.4b)] of the possible states, but it does not capture the dynamical correlations around nodes, since we use a heterogeneous mean-field description [Eq. (6.4a)]. These correlations become especially important in the presence of hubs with a large membership, which is the case for the hypergraph considered in Fig. 6.3. In fact, when we use the same group size distribution as in Fig. 6.3, but with a more homogeneous membership distribution, the discrepancy at high prevalence disappears (see the Supplementary Information).

## 6.6 Phase transition

In this section, we unveil the important role of influential groups on the phase transition of hypergraph contagions. We first derive general expression for critical points, marking the limit of the domain of validity of a stationary solution. Secondly, we obtain an expression for tricritical points, indicating when

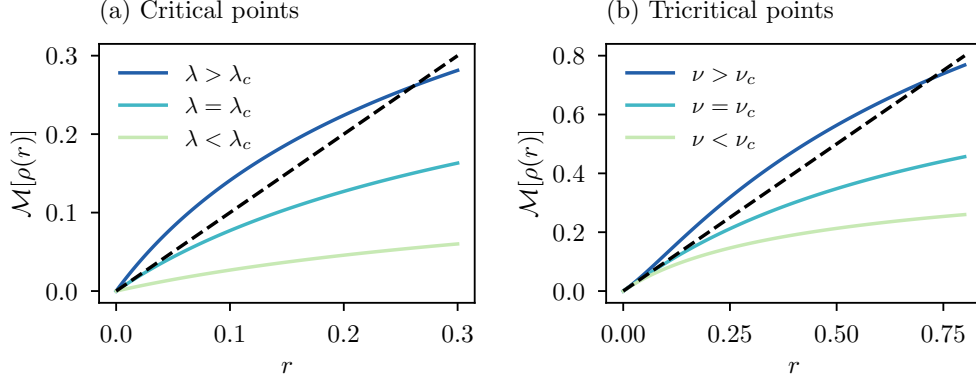


FIGURE 6.4 – Critical behavior of the function  $\mathcal{M}[\rho(r)]$ . Each intersection with the dashed line represents a stationary solution of Eq. (6.7). Results refer to the infection rate function  $\beta(n, i) = \lambda i^\nu$  and a hypergraph with  $g_m = (\delta_{m,2} + \delta_{m,3})/2$  and  $p_n = (\delta_{n,4} + \delta_{n,5})/2$ . (a) We fix  $\nu = 1$  and vary  $\lambda$ . For  $\lambda > \lambda_c$ , a nontrivial solution emerges. (b) We fix  $\lambda = \lambda_c$  and vary  $\nu$ . Note that the slope of all solid lines is 1 at the origin. A nonlinear exponent  $\nu > \nu_c$  is associated with a discontinuous phase transition, since  $\partial_r^2 \mathcal{M} > 0$  in this case.

the phase transition switches from continuous to discontinuous, and a bistable regime appears. These results are valid for any infection rate function such that  $\beta(n, i) > 0$  for all  $i > 0$  (see Supplementary Information for a consideration of threshold models).

We then concentrate our study on the infection rate function  $\beta(n, i) = \lambda i^\nu$ . This allows us to define important threshold values, the *invasion threshold*  $\lambda_c$  above which the disease-free solution  $I^* = 0$  is unstable, and the *bistability threshold*  $\nu_c$ , the smallest nonlinear exponent allowing for a discontinuous phase transition. In particular, we will show how heterogeneous membership and group size distributions alter these thresholds, especially in the presence of a mesoscopic localization driven by influential groups.

In the following, we assume that the system has reached the stationary state and we drop the asterisk to simplify the notation throughout this section.

### 6.6.1 Critical points and the invasion threshold

Equations (6.7a) and (6.8) imply that each  $s_m$  and  $f_{n,i}$  can be written in terms of  $r$  and  $\rho$ . In turn,  $r$  and  $\rho$  can be written in terms of  $s_m$  and  $f_{n,i}$  through Eqs. (6.2) and (6.3), which means the stationary solutions are determined by a set of self-consistent equations.

Satisfying all self-consistent equations means we can reexpress all quantities  $(s_m, f_{n,i}, \rho, r)$  as functions of a single mean-field quantity, which we choose to be  $r$ .  $s_m(r)$  is given by Eq. (6.7a), and  $\rho(r)$  is given by Eq. (6.3), which is rewritten as

$$\rho(r) = r \frac{\sum_m m(m-1)s_m(r)g_m}{\sum_m m s_m(r)g_m}. \quad (6.9)$$

$f_{n,i}$  is more simply written as a composite function,  $f_{n,i}[\rho(r)]$ , given by Eq. (6.8). Finally,  $r$  itself must

satisfy Eq. (6.2), which we can write as  $r = \mathcal{M}[\rho(r)]$  where

$$\mathcal{M}(\rho) = \frac{\sum_{n,i} \beta(n,i)(n-i)f_{n,i}(\rho)p_n}{\sum_{n,i} (n-i)f_{n,i}(\rho)p_n}. \quad (6.10)$$

In Fig. 6.4, the stationary solutions thus correspond to the intersections of  $\mathcal{M}[\rho(r)]$  (solid lines) and  $r$  (dashed line). We see that  $r = 0$  is always a solution, while the solution  $r > 0$  only exists for certain values of the parameter  $\lambda$ . This indicates the presence of a critical point.

Critical points mark the limit of the domain of validity of a solution to the equation  $r = \mathcal{M}[\rho(r)]$ . They arise when  $r$  is tangent to  $\mathcal{M}[\rho(r)]$ , which implies

$$\frac{d\mathcal{M}}{dr} = \frac{d\mathcal{M}}{d\rho} \frac{d\rho}{dr} = 1, \quad (6.11)$$

as can be seen in Fig. 6.4(a), where  $\mathcal{M}[\rho(r)]$  is tangent to  $r$  at the point  $r = 0$  for some value  $\lambda = \lambda_c$ .

When the tangent point  $r > 0$ , Eq. (6.11) needs to be solved numerically. However when the solution  $r \rightarrow 0$ , we are able to obtain an analytical expression. In this limit, we expect to have  $s_m \rightarrow 1$ ,  $\rho \rightarrow 0$  and  $f_{n,i} \rightarrow \delta_{i,0} \forall n$ , i.e., all nodes are susceptible. Therefore, from Eq. (6.9) we have

$$\left. \frac{d\rho}{dr} \right|_{r \rightarrow 0} = \frac{\langle m(m-1) \rangle}{\langle m \rangle}, \quad (6.12)$$

where  $\langle \dots \rangle$  stands for the expectation value with respect either to  $g_m$  or  $p_n$ . From Eq. (6.10) (and using the fact that  $\beta(n,0) \equiv 0$ ) we also obtain

$$\left. \frac{d\mathcal{M}}{d\rho} \right|_{\rho \rightarrow 0} = \frac{1}{\langle n \rangle} \sum_{n,i} \beta(n,i)(n-i)p_n \underbrace{\left. \frac{df_{n,i}}{d\rho} \right|_{\rho \rightarrow 0}}_{h_{n,i}}. \quad (6.13)$$

To evaluate  $h_{n,i}$ , we apply the derivative  $d/d\rho$  to Eq. (6.8). We obtain  $h_{n,1} = n$  and

$$h_{n,i} = \frac{n! \prod_{j=1}^{i-1} \beta(n,j)}{(n-i)! i!} \quad \forall i \in \{2, \dots, n\}, \quad (6.14)$$

Also,  $\sum_i h_{n,i} \equiv 0$  as  $\sum_i f_{n,i} = 1$ , hence

$$h_{n,0} = - \sum_{i>0} h_{n,i}.$$

Combining Eqs (6.12), (6.13) and (6.14) with Eq. (6.11), we obtain the following implicit expression for critical points in the limit  $r \rightarrow 0$

$$\frac{\langle m(m-1) \rangle}{\langle m \rangle \langle n \rangle} \left\langle \sum_{i>0} \frac{n!}{(n-i-1)! i!} \prod_{j=1}^i \beta(n,j) \right\rangle = 1. \quad (6.15)$$

For the infection rate function  $\beta(n,i) = \lambda i^\nu$ , Eq. (6.15) defines the *invasion* threshold  $\lambda_c$ , i.e., the critical value of  $\lambda$  marking the limit of the validity for a solution  $r \rightarrow 0$ . This solution is not always stable, but we always have that the trivial solution  $r = 0$ , corresponding to  $I = 0$ , is unstable for all  $\lambda > \lambda_c$ . This is illustrated in Fig. 6.5(a) and 6.5(b) for regular hypergraph structures with fixed membership and group size. The invasion threshold depends on both the structure  $(g_m, p_n)$  and the

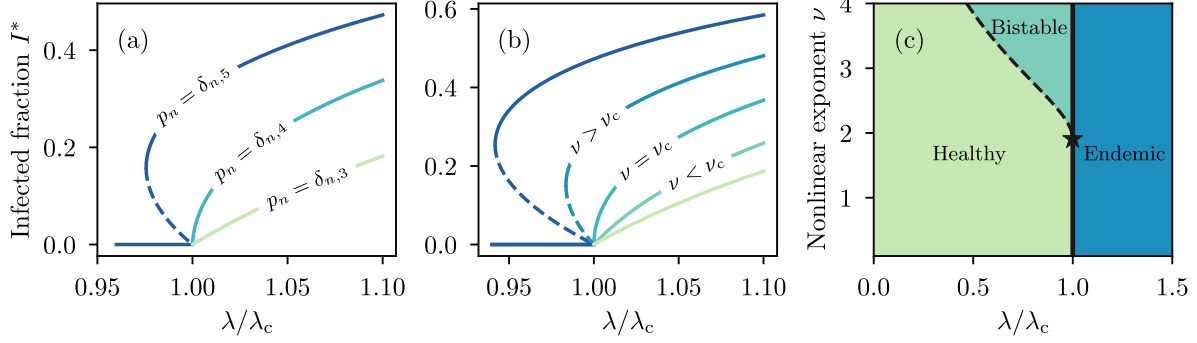


FIGURE 6.5 – Phase transition in regular hypergraphs. The membership and group size distributions are of the form  $(g_m, p_n) = (\delta_{m,m_0}, \delta_{n,n_0})$ . The infection rate function is  $\beta(n, i) = \lambda i^\nu$ . Solid and dashed lines in (a) and (b) represent stable and unstable solutions respectively for the stationary fraction of infected nodes. (a)  $\nu = 1.8$ , but different structures were used, with a node always having the same number of different neighbors. From bottom to top, we have  $(g_m, p_n) = (\delta_{m,6}, \delta_{n,3})$ ,  $(\delta_{m,4}, \delta_{n,4})$  and  $(\delta_{m,3}, \delta_{n,5})$ . Note that for the middle curve  $(\delta_{m,4}, \delta_{n,4})$ ,  $\nu_c \approx 1.81$ . (b) The hypergraph structure is fixed with  $g_m = \delta_{m,3}$  and  $p_n = \delta_{n,4}$ . Values of  $\nu \in \{1.5, 1.7, \nu_c, 2.1, 2.3\}$  (bottom to top curves) were used, with  $\nu_c \approx 1.91$ . (c) Phase diagram for the same hypergraph as in (b). The dashed critical line ( $\lambda = \lambda_p$ ) and the solid critical line ( $\lambda = \lambda_c$ ) coalesce at the tricritical point  $(\lambda_c, \nu_c)$  indicated by the star marker.

nonlinear exponent  $\nu$ . The resulting phase diagram in the  $(\lambda, \nu)$  plane is shown in Fig. 6.5(c). The invasion threshold spans the critical line (solid line)  $\lambda = \lambda_c$  in the phase diagram.

The dashed critical line  $\lambda = \lambda_p$  is associated with the limit of validity of a solution where  $\mathcal{M}[\rho(r)]$  is tangent to  $r$  for some  $r > 0$ —it is thus solved numerically. We call  $\lambda_p$  the *persistence* threshold as it is the smallest value of  $\lambda$  such that a nontrivial solution is locally stable. For continuous phase transitions,  $\lambda_c = \lambda_p$ , but for discontinuous phase transitions,  $\lambda_p < \lambda_c$ .

### 6.6.2 Tricritical points and the bistability threshold

Depending on the structure and the form of  $\beta(n, i)$ , we can have a continuous or a discontinuous phase transition, as can be seen in Fig. 6.5. When the phase transition is continuous, we have possibly two solutions for the stationary fraction of infected nodes,  $I_1 = 0$  and  $I_2 > 0$ . When  $I_2$  exists (for instance when  $\lambda > \lambda_c$ ),  $I_1$  is unstable.

When the phase transition is discontinuous, we have typically three solutions,  $I_1 = 0$  and  $0 < I_2 < I_3$ . In the bistable regime [for instance when  $\lambda \in (\lambda_p, \lambda_c)$ ], all three solutions coexist,  $I_2$  is unstable, and  $I_1$  and  $I_3$  are locally stable. In the endemic regime [for instance when  $\lambda \geq \lambda_c$ ], only  $I_1$  and  $I_3$  exist, and only  $I_3$  is locally stable.

We are interested in knowing when the phase transition changes from continuous to discontinuous. In Fig. 6.5(c), the bistable regime starts at a tricritical point (star marker), where two critical lines meet. For the infection rate function  $\beta(n, i) = \lambda i^\nu$  and a fixed hypergraph structure, the tricritical point happens at  $(\lambda, \nu) = (\lambda_c, \nu_c)$ , where  $\nu_c$  is what we call the *bistability* threshold, since a bistable regime only exists for  $\nu > \nu_c$ .

To get some insights on the properties of tricritical points, we show in Fig. 6.4(b) the function  $\mathcal{M}[\rho(r)]$

at  $\lambda = \lambda_c$  and for values of  $\nu$  below, at and above the bistability threshold. For  $\nu < \nu_c$ , we have  $\partial_r^2 \mathcal{M}|_{r \rightarrow 0} < 0$  and  $I_3$  does not exist. For  $\nu > \nu_c$ , we have  $\partial_r^2 \mathcal{M}|_{r \rightarrow 0} > 0$  and there exists a solution  $I_3 > 0$ . At the tricritical point,  $I_3 = I_2 \rightarrow 0$ , hence the nontrivial solution is degenerate, which is possible only if

$$\left. \frac{d^2 \mathcal{M}}{dr^2} \right|_{r \rightarrow 0} = \left[ \frac{d\mathcal{M}}{d\rho} \frac{d^2 \rho}{dr^2} + \left( \frac{d\rho}{dr} \right)^2 \frac{d^2 \mathcal{M}}{d\rho^2} \right] \Big|_{r \rightarrow 0} = 0 .$$

Since a tricritical point is also a critical point, from Eq. (6.11)  $d\mathcal{M}/d\rho = (d\rho/dr)^{-1}$ , so the condition can be rewritten as

$$\left[ \frac{d^2 \rho}{dr^2} + \left( \frac{d\rho}{dr} \right)^3 \frac{d^2 \mathcal{M}}{d\rho^2} \right] \Big|_{r \rightarrow 0} = 0 .$$

The derivatives on  $\rho$  with respect to  $r$  at a critical point where  $r \rightarrow 0$  can be easily evaluated, and the condition now becomes

$$2 \left( \frac{\langle m^2 \rangle^2}{\langle m \rangle^2} - \frac{\langle m^3 \rangle}{\langle m \rangle} \right) + \left( \frac{\langle m(m-1) \rangle}{\langle m \rangle} \right)^3 \left. \frac{d^2 \mathcal{M}}{d\rho^2} \right|_{\rho \rightarrow 0} = 0 . \quad (6.16)$$

To evaluate the last term of Eq. (6.16), let us rewrite

$$\mathcal{M}(\rho) = \frac{\sum_{n,i} \beta(n,i)(n-i)f_{n,i}(\rho)p_n}{\sum_{n,i} (n-i)f_{n,i}(\rho)p_n} \equiv \frac{u(\rho)}{v(\rho)} .$$

In the limit  $\rho \rightarrow 0$ ,  $f_{n,i} \rightarrow \delta_{i,0}$ , which implies  $u(\rho) \rightarrow 0$  and  $v(\rho) \rightarrow \langle n \rangle$ , therefore

$$\left. \frac{d^2 \mathcal{M}}{d\rho^2} \right|_{\rho \rightarrow 0} = \frac{1}{\langle n \rangle} \left. \frac{d^2 u}{d\rho^2} \right|_{\rho \rightarrow 0} - \frac{2}{\langle n \rangle^2} \left. \frac{du}{d\rho} \frac{dv}{d\rho} \right|_{\rho \rightarrow 0} . \quad (6.17)$$

First order derivatives can be evaluated using

$$\left. \frac{du}{d\rho} \right|_{\rho \rightarrow 0} = \sum_{n,i} \beta(n,i)(n-i)p_n h_{n,i} , \quad (6.18a)$$

$$\left. \frac{dv}{d\rho} \right|_{\rho \rightarrow 0} = \sum_{n,i} (n-i)p_n h_{n,i} . \quad (6.18b)$$

For the second-order derivative, let us define  $l_{n,i} \equiv d^2 f_{n,i}/d\rho^2|_{\rho \rightarrow 0}$ , so that we can write

$$\left. \frac{d^2 u}{d\rho^2} \right|_{r \rightarrow 0} = \sum_{n,i} \beta(n,i)(n-i)p_n l_{n,i} . \quad (6.19)$$

Finally, we can apply the second order derivative to Eq. (6.7b) to obtain the recurrence relation

$$\begin{aligned} (i+1)l_{n,i+1} &= 2[(n-i)h_{n,i} - (n-i+1)h_{n,i-1}] , \\ &+ [i + (n-i)\beta(n,i)] l_{n,i} , \\ &- (n-i+1)\beta(n,i-1)l_{n,i-1} . \end{aligned} \quad (6.20)$$

Again,  $l_{n,0} = -\sum_{i>0} l_{n,i}$  by definition.

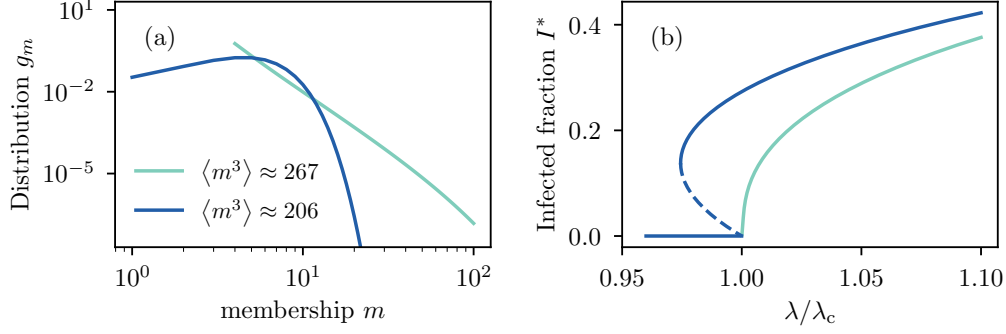


FIGURE 6.6 – Impact of the third moment of the membership distribution on the phase transition. The group size distribution is fixed to  $p_n = \delta_{n,3}$ . (a) Membership distributions with the same first two moments ( $\langle m \rangle \approx 5.03$  and  $\langle m^2 \rangle \approx 30.2$ ) but with a different third moment  $\langle m^3 \rangle$ . The distribution with higher  $\langle m^3 \rangle$  is of the form  $g_m \propto (a+m)^{-b} e^{-m/c}$  with  $(a, b, c) = (-0.8, 3.72, 40)$  and ranges from  $m_{\min} = 4$  to  $m_{\max} = 100$ . The one with lower  $\langle m^3 \rangle$  is Poisson distributed  $g_m \propto a^m e^{-a}/m!$  with  $a = 5$  and ranges from  $m_{\min} = 1$  to  $m_{\max} = 100$ . (b) Non-trivial solutions to Eq. (6.7) for the corresponding membership distribution. Solid and dashed lines represent stable and unstable solutions, respectively. The infection rate function considered is  $\beta(n, i) = \lambda i^\nu$  with  $\nu = 2.8$ .

Even though it is possible to express the  $l_{n,i}$  in an explicit form, the expression does not give us more intuition, and it is simpler to calculate the  $l_{n,i}$  using the recurrence equation just given. After rewriting  $d^2 \mathcal{M}/d\rho^2|_{\rho \rightarrow 0} \equiv F[p_n, \beta(n, i)]$ , tricritical points are obtained by solving the equation

$$\frac{\langle m^2 \rangle^2}{\langle m \rangle^2} - \frac{\langle m^3 \rangle}{\langle m \rangle} + \frac{\langle m(m-1) \rangle^3}{2 \langle m \rangle^3} F[p_n, \beta(n, i)] = 0. \quad (6.21)$$

Tricritical points result from an intricate relation between the structure  $(g_m, p_n)$  and the infection rate  $\beta(n, i)$ . Fig. 6.5 shows that either changing the structure [Fig. 6.5(a)] or the shape of the infection rate function [Fig. 6.5(b)-(c)] can lead to a change of behavior, from a continuous phase transition to a discontinuous one with a bistable regime.

A first hypothesis we can make from these simple examples is that more nonlinear infection rates (larger  $\nu$ ) and larger groups promote bistability. However, we will see that this intuition does not hold in general for heterogeneous structures due to the onset of mesoscopic localization.

### 6.6.3 Heterogeneous memberships

In this section we investigate the effects of a heterogeneous membership distribution  $g_m$  while keeping  $p_n = \delta_{n,n_0}$  homogeneous to disentangle the impact of the different types of heterogeneity. A first remark we can make about the invasion threshold [Eq. (6.15)] is that it is coherent with heterogeneous pair-approximation frameworks [193] on random networks when only dyadic interactions are considered, i.e., when  $p_n = \delta_{n,2}$ . In this case, we can set  $\nu = 1$  without loss of generality, thus recovering the standard SIS model. The associated threshold is

$$\lambda_c^{\text{SIS}} = \frac{\langle m \rangle}{\langle m(m-1) \rangle},$$

where  $g_m$  can now be interpreted as the standard degree distribution of graphs. This threshold, although quite accurate for most structures, does not capture the hub reinfection mechanism [91, 268], and thus could be inaccurate for graphs with hubs of a very large degree.

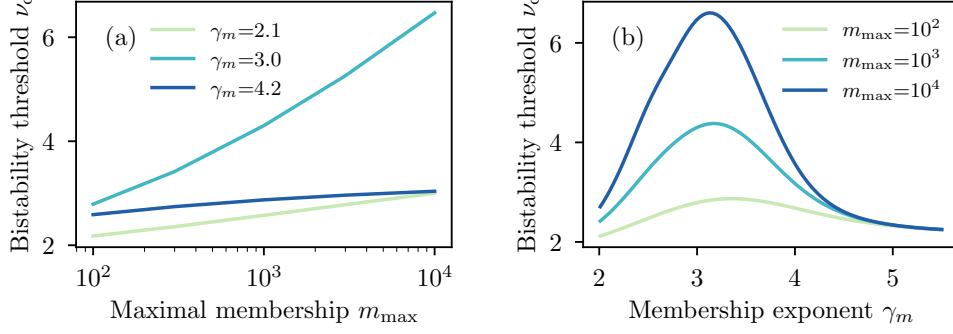


FIGURE 6.7 – Impact of heterogeneous memberships on the bistability threshold. We considered hypergraphs with power-law membership distributions  $g_m \propto m^{-\gamma_m}$  with various exponents  $\gamma_m$ , different maximal values  $m_{\max}$ , a minimal value  $m_{\min} = 2$ , and a regular group size distribution  $p_n = \delta_{n,4}$ . We obtain the bistability threshold by solving Eq. (6.21). Lower values of  $\gamma_m$  imply a more heterogeneous membership distribution. (a) The bistability threshold grows logarithmically or faster for  $\gamma_m \leq 4$ , but converges for  $\gamma_m > 4$ . (b) The bistability threshold shows a maximum near  $\gamma_m = 3$ , suggesting that this is the maximum in the limit  $m_{\max} \rightarrow \infty$ .

More generally, for group interactions ( $p_n \neq \delta_{n,2}$ ) we can see that a larger average excess membership  $\langle m(m-1) \rangle / \langle m \rangle$  always leads to a smaller invasion threshold  $\lambda_c$ , akin to the standard SIS model, but the relationship is now nonlinear. To see this, let us rewrite Eq. (6.15) as

$$\frac{1}{\langle n \rangle} \left\langle \sum_{i>0} \frac{n!}{(n-i-1)! i!} \prod_{j=1}^i \beta(n, j) \right\rangle = \frac{\langle m \rangle}{\langle m(m-1) \rangle}. \quad (6.22)$$

Since  $\beta(n, i)$  is a monotonically increasing function of  $\lambda$  for all  $n, i$ , then the left-hand side of Eq. (6.22) is a monotonically increasing function (of  $\lambda$ ) as well. Consequently, if the right-hand side decreases,  $\lambda_c$  must decrease as well.

Assessing the impact of membership heterogeneity on the bistability threshold  $\nu_c$  is more complicated. In fact, Eq. (6.21) explicitly depends on the first three moments of  $g_m$ , but it also depends on the first two moments implicitly through  $\lambda_c$ , at which  $F$  must be evaluated.

In order to build our intuition, let us assume that we are able to keep fixed the first two moments  $\langle m \rangle$  and  $\langle m^2 \rangle$  while increasing  $\langle m^3 \rangle$ . This means that  $\lambda_c$  would not change, hence the only dependence on  $g_m$  would be explicit in Eq. (6.21). Since the term depending on  $\langle m^3 \rangle$  is negative, increasing the third moment implies that  $F$  must increase if we want to balance Eq. (6.21). But since  $d^2\mathcal{M}/dr^2|_{r \rightarrow 0}$  increases with  $\nu$  [see Fig. 6.4], and thereby  $F$  as well, we can conclude that increasing  $\langle m^3 \rangle$  while keeping the first two moments fixed leads to an increase of the bistability threshold  $\nu_c$ . This is validated in Fig. 6.6, where we consider two distributions  $g_m$  sharing the same first two moments, but a different third moment [see Fig. 6.6(a)]. By comparing the associated epidemic curves in Fig. 6.6(b), it is evident how the larger third moment suppresses the emergence of a bistable regime.

A corollary of this argument is that for certain structures, it is impossible to have bistability. To see this, let us consider a power-law membership distribution of the form  $g_m \propto m^{-\gamma_m}$ . In this case, since the bistability threshold depends on the third moment of  $g_m$ , while the invasion threshold only depends on the first two, by setting the exponent  $3 < \gamma_m < 4$ , the invasion threshold converges to a value  $\lambda_c > 0$ , but  $\nu_c$  does not exist. In other words, it is impossible to have a discontinuous phase

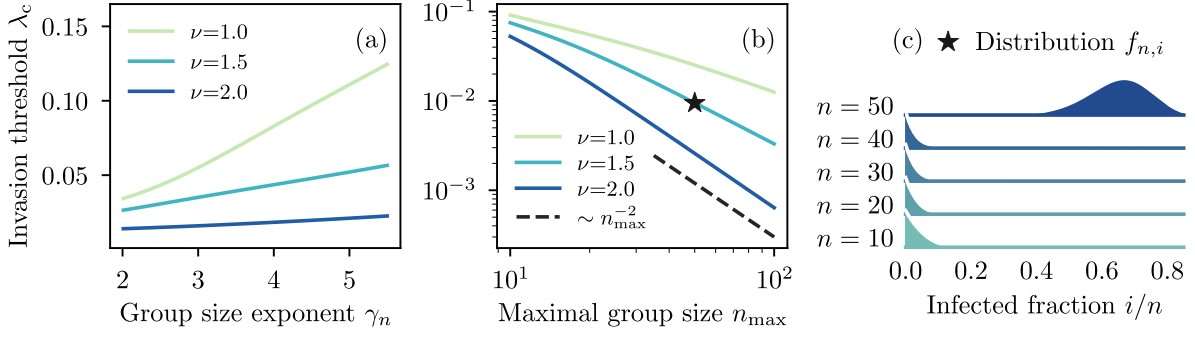


FIGURE 6.8 – Impact of heterogeneous group sizes on the invasion threshold. We considered hypergraphs with power-law group size distributions  $p_n \propto n^{-\gamma_n}$  with various exponents  $\gamma_n$ , a regular membership distribution  $g_m = \delta_{m,4}$ , and various exponents  $\nu$  for the infection function in Eq. (6.1). We obtain the invasion threshold using Eq. (6.15). Lower values of  $\gamma_n$  implies a more heterogeneous group size distribution. (a) For a fixed  $n_{\max} = 20$ , the invasion threshold increases with larger  $\gamma_n$ , but the effect is more limited for larger  $\nu$ . (b) For a fixed  $\gamma_n = 3$ , the invasion threshold decreases like  $n_{\max}^{-\nu}$  for large  $n_{\max}$ , indicating the onset of mesoscopic localization (see chapter 3 and 4). (c) Stationary distribution  $f_{n,i}$  for the starred case in (b), i.e.,  $\gamma_n = 3$  and  $\nu = 1.5$ , with  $\lambda = 1.1\lambda_c$ , illustrating localization in the largest groups.

transition.

This second statement is validated in 6.7(a), where we show that that the bistability threshold  $\nu_c$  appears to grow without bound as  $m_{\max} \rightarrow \infty$  for  $\gamma_m \leq 4$ . Instead, for  $\gamma_m \geq 4$ , the bistability threshold appears to converge, as expected, since the first three moments of  $g_m$  converge as well. What is more surprising is the nonmonotonic behavior of  $\nu_c$  with respect to  $\gamma_m$ , which we present in Fig. 6.7(b). The bistability threshold has a well defined maximum at a value of  $\gamma_m$  that appears to converge to  $\gamma_m = 3$  for  $m_{\max} \rightarrow \infty$ . In other words,  $\gamma_m = 3$  is the optimal value of membership exponent in suppressing the emergence of a discontinuous phase transition and the related bistability.

This can be understood from Eq. (6.21) : for  $\gamma_m > 3$ , the invasion threshold does not vary much since the first two moments of  $g_m$  are finite. Hence maximizing the third moment maximizes  $\nu_c$ , which corresponds to  $\gamma_m \rightarrow 3$ . One could still be surprised that the bistability threshold grows more slowly with  $m_{\max}$  in the range  $2 < \gamma_m < 3$ , since the invasion threshold  $\lambda_c$  tends toward zero. In this case, the bistable regime exists, but its width  $(\lambda_p, \lambda_c)$  simply vanishes as  $\lambda_c \rightarrow 0$ .

#### 6.6.4 Heterogeneous group sizes

Let us now consider hypergraphs with heterogeneous group size distribution  $p_n$ , and homogeneous membership distribution, namely  $g_m = \delta_{m,m_0}$ . In this case, the invasion threshold, as defined by Eq. (6.15), depends on the whole distribution  $p_n$ , which makes drawing general conclusions on the impact of a heterogeneous distribution  $p_n$  more difficult.

To get some intuitions, let us consider the standard SIS model, i.e., the case  $\nu = 1$  in Eq. (6.1). With our AMEs, it was shown that [266]

$$\lambda_c^{-1} \sim \left( \frac{\langle m(m-1) \rangle}{\langle m \rangle} \right) \left( \frac{\langle n(n-1) \rangle}{\langle n \rangle} \right) + n_{\max}, \quad (6.23)$$



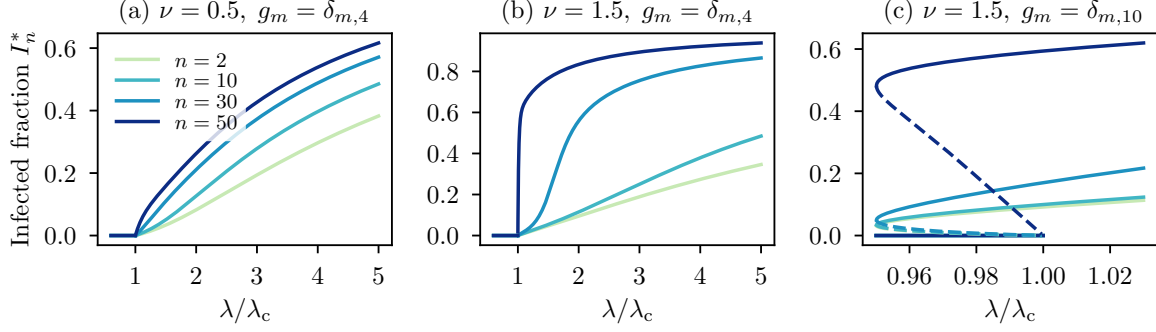


FIGURE 6.9 – Mesoscopic localization in large groups, illustrated by  $I_n^*$ , the average stationary fraction of infected nodes in groups of various size  $n$ . We considered hypergraphs with power-law group size distributions  $p_n \propto n^{-\gamma_n}$  with  $\gamma_n = 3$ , regular membership distributions of the form  $g_m = \delta_{m,m_0}$ , and various exponents  $\nu$  for the infection function in Eq. (6.1). Near the invasion threshold  $\lambda_c$ ,  $I_n^*$  is larger for larger groups in (a) with  $\nu = 0.5$ , but localization is much more pronounced for (b) with  $\nu = 1.5$ . (c) For discontinuous phase transitions, mesoscopic localization is still possible, but near  $\lambda_c$  we must look at the unstable solution for  $I_n^*$  in the bistable regime.

for power-law distributions  $p_n \propto n^{-\gamma_n}$  with large cut-offs  $n_{\max}$ . The first term on the right-hand side of Eq. (6.23) suggests that more heterogeneous groups-size distributions  $p_n$  (smaller values of  $\gamma_n$ ) lead to smaller invasion thresholds. Intuition tells us that we should expect this behavior for  $\nu \neq 1$  as well. We have therefore investigated numerically in Fig. 6.8(a) the invasion threshold as a function of the group size exponent for different values of  $\nu$ , confirming that more heterogeneous group sizes (smaller  $\gamma_n$ ) do lead to a smaller invasion threshold, even for nonlinear infection functions ( $\nu \neq 1$ ). However, this effect is mitigated when larger values of  $\nu$  are considered. For large  $\nu$  and large  $n_{\max}$ , the value of the invasion threshold is dominated by the cut-off, and scales as  $\lambda_c \sim n_{\max}^{-\nu}$ , as illustrated in Fig. 6.8(b).

This behavior can be attributed to the onset of *mesoscopic localization*. In chapter 3 and 4, it was shown analytically for  $\nu = 1$  that, for certain combinations of  $(\gamma_m, \gamma_n)$ , the epidemic near the invasion threshold [ $\lambda = \lambda_c(1 + \epsilon)$  with  $\epsilon \ll 1$ ] is dominated by the largest most influential groups. In these scenarios, the second term on the right-hand side in Eq. (6.23) dominates the first one and, near  $\lambda_c$ , the group prevalence  $I_n$  grows exponentially with  $n$ , i.e.,  $I_n \sim n^{1/2} e^{an}$  for some constant  $a > 0$ . While an analytical characterization of mesoscopic localization in the general case of  $\nu \neq 1$  is out of the scope of this chapter, we provide clear numerical evidence of localization phenomena in Fig. 6.8(c). The stationary distributions of the fraction of infected nodes in groups of increasing size  $n$  is concentrated in the largest group ( $n = 50$ ) near the invasion threshold  $\lambda_c$ .

Since mesoscopic localization was characterized using a linear contagion ( $\nu = 1$ ) and a continuous phase transition in chapter 3 and 4, two natural questions arise : How does  $\nu \neq 1$  affects localization ? And what happens in the context of discontinuous phase transitions ? In Fig. 6.9, we present the phase diagram of the group prevalence  $I_n$  for different scenarios. Comparing Fig. 6.9(a) and Fig. 6.9(b), we see that increasing  $\nu$  from 0.5 to 1.5 (while keeping  $g_m = \delta_{m,4}$ ) strengthens localization effects, which is expected since reinforcement effects are more important when the group prevalence is high. In Fig. 6.9(c), we show a similar diagram, but for a discontinuous phase transition. We see that the concentration of infected nodes in the largest groups is still possible, but the phenomenon is now associated with the *unstable* solution near the invasion threshold [ $\lambda = \lambda_c(1 - \epsilon)$  with  $\epsilon \ll 1$ ]. Therefore,

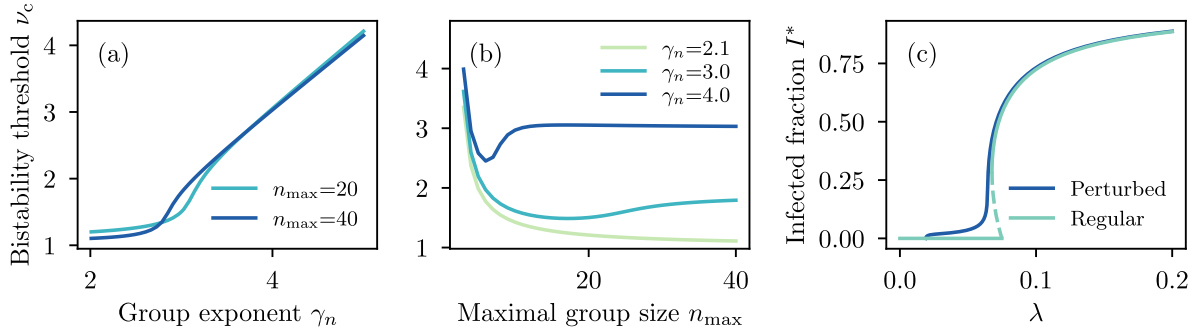


FIGURE 6.10 – Impact of heterogeneous group sizes on the bistability threshold. We considered hypergraphs with power-law group size distributions  $p_n \propto n^{-\gamma_n}$  with various exponents  $\gamma_n$  and a regular membership distribution  $g_m = \delta_{m,4}$ . We solve the bistability threshold using Eq. (6.21). Lower values of  $\gamma_n$  implies a more heterogeneous group size distribution. (a) For a fixed  $n_{\max}$ , the bistability threshold increases with larger  $\gamma_n$ . (b) For a fixed  $\gamma_n$ , the bistability threshold has a nonmonotonic relationship with  $n_{\max}$ . (c) Phase transition using  $g_m = \delta_{m,4}$  and  $\nu = 2.3$ . We use  $p_n = \delta_{n,4}$  for the regular case, and  $p_n = (1-\epsilon)\delta_{n,4} + \epsilon\delta_{n,15}$  with  $\epsilon = 10^{-3}$  for the perturbed case. Mesoscopic localization (see chapter 3 and 4) inhibits bistability in the perturbed case.

mesoscopic localization affects both continuous and discontinuous phase transition with a bistable regime, but the exponential growth of  $I_n$  with  $n$  near  $\lambda_c$  concerns the stable solution in the former and the unstable solution in the latter.

If we now reinterpret the results of Fig. 6.8 in light of these considerations, larger values of  $\nu$  facilitate the onset of mesoscopic localization, where the largest groups drive the onset of the endemic phase, and make the invasion threshold scale as  $\lambda_c \sim n_{\max}^{-\nu}$ . This explains why  $\lambda_c$  varies only slightly with  $\gamma_n$  for  $\nu = 2$  in Fig. 6.8(a).

In Fig. 6.10, we finally investigate the role of heterogeneous group sizes on the bistability threshold by varying the group exponent  $\gamma_n$  and the maximal group size  $n_{\max}$ . From Fig. 6.10(a), we see that a more heterogeneous group distribution, thereby increasing the fraction of larger groups, decreases the value of the bistability threshold  $\nu_c$ . This is consistent with our observation on regular structures [Fig. 6.5], for which larger groups appear to promote bistability. However, Fig. 6.10(b) brings some nuance to this statement : for a fixed exponent  $\gamma_n$ , there is an interesting nonmonotonic relationship between  $\nu_c$  and the largest group  $n_{\max}$ . As such, the presence of larger groups does not always promote bistability.

We can again attribute this behavior to localization effects. In fact, we are able to illustrate this via a very simple example in Fig. 6.10(c). We look at the phase transition for a regular hypergraph with a fixed group size,  $p_n = \delta_{n,4}$ , and a perturbed version of it, where we introduce a small proportion of larger groups,  $p_n = (1-\epsilon)\delta_{n,4} + \epsilon\delta_{n,15}$  with  $\epsilon = 10^{-3}$ . For the regular distribution, the phase transition is discontinuous, while for the perturbed distribution it is continuous, with the contagion localized in the largest groups near the invasion threshold. The bistability threshold  $\nu_c$  is larger for the perturbed distribution since mesoscopic localization reduces considerably the invasion threshold  $\lambda_c$ . The largest most influential groups drive and self-sustain an endemic state for smaller values of  $\lambda$ , hence preventing a bistable regime.

## 6.7 Influence maximization

Influence maximization broadly refers to the problem of selecting a subset of nodes to initially spark a diffusion process in order to maximize the effect. The process could represent the spread of information, the diffusion of innovations or a viral marketing campaign [76, 244].

There is a large body of literature on influence maximization in complex networks, where various models have been used : threshold models [54, 73, 158, 211], independent cascade [158], and simple contagion models (SI, SIS, SIR) [53, 83, 164, 239], to name a few. Recently, these ideas have been also exported to higher-order networks [10, 305].

The effectiveness of an influence maximization procedure is often measured by the fraction of affected nodes (in the limit  $t \rightarrow \infty$ ) for processes that terminate. However, because the final epidemic size in the SIS dynamics does not depend on the seeds (other than for stochastic extinction), we will consider the simpler task of maximizing  $\dot{I}(0)$ , the initial spreading speed. This is often a straightforward task to solve for graphs. Considering the SIR model for instance, one just needs to maximize the number of outgoing edges from infected to susceptible nodes, which implies that nodes of maximal degree would be optimal influencers. However, we will show that additional considerations need to be accounted for in higher-order networks. More specifically, our goal is to use our formalism to answer the following question : *Should we focus on finding influential nodes, or seed the spread from influential groups ?*

In this section, to simplify the notation, all dynamic quantities are evaluated at  $t = 0$ , e.g.  $I(0) \equiv I$ .

Let us assume that we are given a fixed hypergraph and an initial fraction of nodes that can be infected at the initial time  $I = \epsilon \ll 1$  (the seeds of the contagion). Our task is to invade the system as fast as possible by maximizing  $\dot{I}$  for a hypergraph contagion, which is equivalent to maximizing the objective function

$$\Phi[\mathcal{S}, \mathcal{F}] = r \langle ms_m \rangle , \quad (6.24)$$

where we define the initial node states  $\mathcal{S} \equiv \{s_m\}_{m=1}^{m_{\max}}$  and the initial group states  $\mathcal{F} \equiv \{f_{n,i} | 0 \leq i \leq n\}_{n=2}^{n_{\max}}$ . The optimization problem is also constrained by

$$0 \leq s_m \leq 1 \quad \forall m , \quad (6.25a)$$

$$0 \leq f_{n,i} \leq 1 \quad \forall n, i , \quad (6.25b)$$

$$\sum_m s_m g_m = 1 - \epsilon , \quad (6.25c)$$

$$\sum_i f_{n,i} = 1 \quad \forall n , \quad (6.25d)$$

$$\frac{\langle ms_m \rangle}{\langle m \rangle} = \frac{1}{\langle n \rangle} \sum_{n,i} (n-i) f_{n,i} p_n . \quad (6.25e)$$

While the first four constraints come from the definitions of the variables, the last one is less straightforward. Equation (6.25e) ensures the consistency between  $\mathcal{S}$  and  $\mathcal{F}$ , more specifically that the fraction of all memberships stubs belonging to susceptible nodes [left-hand side of Eq. (6.25e)] matches the fraction of susceptible nodes in groups [right-hand side of Eq. (6.25e)].

By combining the constraint of Eq. (6.25e) with the definition of  $r$  as given by Eq. (6.2), the objective

function can be simplified as

$$\Phi[\mathcal{S}, \mathcal{F}] \propto \sum_{n,i} \beta(n,i)(n-i)f_{n,i}p_n. \quad (6.26)$$

Although it appears to be independent of  $\mathcal{S}$ , it depends on it implicitly through Eq. (6.25e).

It is worth stressing that our formalism assumes that the membership stubs of nodes are assigned to groups uniformly at random, and thus we cannot *engineer* both  $\mathcal{S}$  and  $\mathcal{F}$ , i.e., choose at the same time the seeds according to their membership and the repartition of the seeds among the various group sizes. Indeed, if we decide for instance to infect only nodes of a certain membership  $m'$  and we try to engineer  $\mathcal{F}$ , there are no guarantees we can achieve such configuration in practice—e.g. we cannot infect a node in a group if none of its nodes have membership  $m'$ .

We therefore compare two strategies to optimize the early spread :

- A) The *influential spreaders* strategy : we engineer  $\mathcal{S}$ , i.e., we choose the fraction of seeds to assign to each membership class, and we assume a random configuration for the groups, i.e., all  $\{f_{n,i}\}_{i=0}^n$  are binomial distributions with probability  $q$  (to be determined).
- B) The *influential groups* strategy : we engineer  $\mathcal{F}$ , i.e., we assign a certain number of seeds in the groups depending on their sizes, and assume that nodes are infected at random through the group to which they belong.

### 6.7.1 Influential spreaders

In this strategy, we are free to engineer  $\mathcal{S}$  in order to maximize  $\Phi$ , with respect to the constraints of Eq. (6.25). Let us assume that  $f_{n,i}$  is a binomial distribution,

$$f_{n,i} = \binom{n}{i} q^i (1-q)^{n-i}.$$

Using Eq. (6.25e), we can identify

$$q = 1 - \frac{\langle ms_m \rangle}{\langle m \rangle}.$$

An optimal solution  $\mathcal{S}^*$  can be found by first finding the value  $q^*$  that maximizes the objective function Eq. (6.26), and then identifying any set  $\mathcal{S}$  that satisfies the relation for  $q = q^*$  above.

There are in general many optimal solutions possible, but they collapse into a single one when  $q$  is sufficiently small, which is reasonable for  $\epsilon \ll 1$ . In this case, we simply have that  $\Phi \approx q$ , and the optimal solution is intuitive : one needs to infect nodes of maximal membership first in order to maximize  $q$ . Interestingly, this is true irrespective of  $\beta(n,i)$ ,  $p_n$  and  $g_m$ .

The infection function and the structure affect the maximal value of  $\epsilon$  such that this solution is unique and optimal. For example, in the simplest case of linear contagion, where  $\beta(n,i) \propto i$ , it is possible to show that this strategy is optimal up to  $q = 1/2$  for all  $g_m$  and  $p_n$ , and we expect even higher values for  $\nu > 1$ . For all practical purposes, targeting nodes of highest membership is optimal, and this is the case in all experiments we considered (see Sec. 6.7.3).

### 6.7.2 Influential groups

In this second strategy, we want to engineer  $\mathcal{F}$  in order to maximize  $\Phi$  with respect to the constraints Eq. (6.25). Let us assume that we can do so by choosing a certain number of groups and infecting a certain portion of their nodes. Following this procedure, one can realize that not all sets  $\mathcal{F}$  satisfying Eq. (6.25) are allowed. For instance, if we decide to infect  $i$  nodes in all groups of size  $n$ , the outcome is different from just having  $f_{n,i} = 1$ . Indeed, nodes belong to more than one group, hence we need to account for spillover effects—groups of size  $n' \neq n$  would have some infected nodes as well, and more than  $i$  nodes could be infected in some groups of size  $n$ .

To do so, let us first define  $\tilde{f}_{n,i}$  as the fraction of all the groups of size  $n$  for which we infect  $i$  nodes at random. Note that if a node belongs to multiple groups, it can be *chosen* more than once for infection, but the duplicates have no effect. Spillovers are taken into account by considering that each of the  $n - i$  nodes that have *not* been chosen for infection in a group of size  $n$  could have been infected in another group, with probability  $u$  (to be determined). Therefore, we can write

$$f_{n,i} = \sum_{j=0}^i \tilde{f}_{n,i-j} B_{n-i+j,j}, \quad (6.27)$$

where

$$B_{k,j} = \binom{k}{j} u^j (1-u)^{k-j}.$$

Second, let us define  $\eta$  as the fraction of all spots in groups that have been chosen for infection,

$$\eta \equiv \frac{1}{\langle n \rangle} \sum_{n,i} i \tilde{f}_{n,i} p_n. \quad (6.28)$$

Since nodes within groups are chosen at random, a node of membership  $m$  is susceptible if it has not been chosen for infection in any of the groups to which it belongs, i.e.,

$$s_m = (1 - \eta)^m.$$

As a consequence,  $\eta$  is constrained by Eq. (6.25c),

$$\langle (1 - \eta)^m \rangle = 1 - \epsilon.$$

The probability  $u$  still needs to be obtained. It corresponds to the fraction of all memberships that are not matched with a spot chosen for infection in a group, but that are still associated with an infected node :

$$u = \frac{\langle m(1 - s_m) \rangle - \eta \langle m \rangle}{(1 - \eta) \langle m \rangle}.$$

With this formulation, we engineer  $\mathcal{F}$  indirectly through  $\tilde{\mathcal{F}} = \{\tilde{f}_{n,i} | 0 \leq i \leq n\}_{n=2}^{n_{\max}}$ . The objective function can be rewritten

$$\begin{aligned} \Phi &\propto \sum_{n,i} \sum_{j=0}^i \beta(n,i) (n-i) \tilde{f}_{n,i-j} B_{n-i+j,j} p_n, \\ &= \sum_{n,i} \sum_{j=0}^{n-i} \beta(n,i+j) (n-i-j) \tilde{f}_{n,i} B_{n-i,j} p_n. \end{aligned}$$

Since the objective function is a linear function of each  $\tilde{f}_{n,i}$ , the optimization problem can be solved using linear programming.

However, there is an intuitive and more efficient way to solve this problem exactly. We just need to identify the most *cost-effective*  $\tilde{f}_{n,i}$  by looking at the effect on  $\Phi$  of increasing  $\tilde{f}_{n,i}$ ,

$$E \equiv \frac{\partial \Phi}{\partial \tilde{f}_{n,i}} \propto \sum_{j=0}^{n-i} \beta(n, i+j)(n-i-j)B_{n-i,j} p_n,$$

versus the *cost* of increasing  $\tilde{f}_{n,i}$ , i.e., the variation of  $\eta$

$$C \equiv \frac{\partial \eta}{\partial \tilde{f}_{n,i}} \propto i p_n.$$

The most cost-effective  $\tilde{f}_{n,i}$  maximizes the ratio

$$R(n, i) = \frac{E}{C} = \frac{1}{i} \sum_{j=0}^{n-i} \beta(n, i+j)(n-i-j)B_{n-i,j}. \quad (6.29)$$

Obviously,  $i = 0$  is always the most cost-effective for all  $n$  (since it has zero cost), but to satisfy Eq. (6.28), we must also fill some  $\tilde{f}_{n,i}$  with  $i > 0$ .

Optimal solutions tend to fill the  $\tilde{f}_{n,i}$  with  $i > 0$  that maximizes  $R(n, i)$ , especially for sufficiently small  $\epsilon$ . A general solution can be obtained using algorithm 1 presented in Methods Sec. 6.10, building on this idea of cost effectiveness. In the worst case, the computational complexity to obtain an optimal solution  $\mathcal{F}^*$  under the influential groups strategy is  $O(m_{\max} + n_{\max}^3)$  when using the method presented in Methods Sec. 6.10, which is much more efficient than using a general-purpose linear-programming method.

Equation (6.29) also gives us an intuition of what defines influential groups when trying to maximize the early spread. If  $\epsilon \ll 1$ , then  $u \ll 1$ , hence we have

$$R(n, i) \approx \frac{\beta(n, i)(n-i)}{i}, \\ \propto i^{\nu-1}(n-i),$$

when considering  $\beta(n, i) = \lambda i^\nu$ . For simple contagions ( $\nu = 1$ ), picking the largest group with a single seed ( $i = 1$ ) is always optimal. For hypergraph contagions with  $\nu > 1$ , the largest groups are the most influential as well, but the optimal number of seeds is generally  $i > 1$ . Hence, beyond its size, the initial configuration of a group determines whether or not it is influential.

### 6.7.3 Experiments

To compare the influential spreaders and the influential groups strategies, we measure the ratio

$$\zeta \equiv \frac{\Phi_{\mathcal{F}}^*}{\Phi_{\mathcal{S}}^*}, \quad (6.30)$$

where  $\Phi_{\mathcal{F}}^*$  and  $\Phi_{\mathcal{S}}^*$  are the values of the objective function for the optimal solution of the influential groups and influential spreader strategies, respectively. Therefore,  $\zeta > 1$  indicates that the influential groups strategy is better to maximize  $\dot{I}$ , and vice versa if  $\zeta < 1$ .

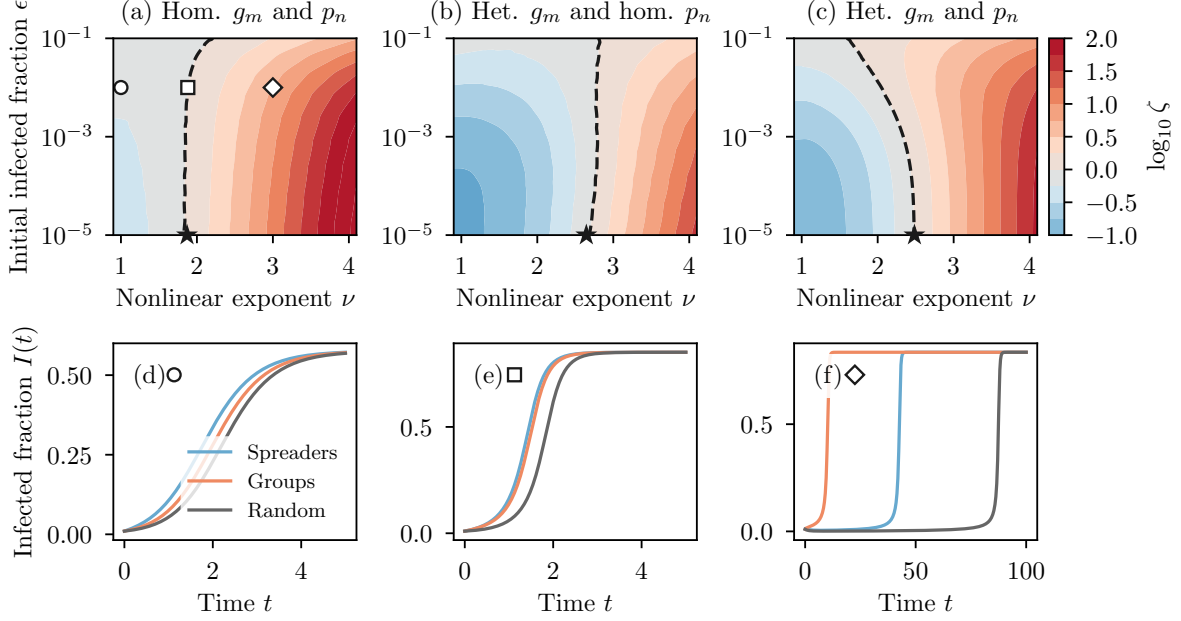


FIGURE 6.11 – Comparison of the influential spreaders and influential groups strategies. (a)-(c) If  $\log_{10} \zeta > 0$ , this indicates that the influential groups strategy is better to maximize  $\dot{I}$ , and vice versa if  $\log_{10} \zeta < 0$ . We use different combinations of homogeneous and heterogeneous distributions. More specifically, for homogeneous distributions we use  $g_m \propto a^m e^{-a}/m!$  and  $p_n \propto a^n e^{-a}/n!$ , with  $a = 5$ ,  $m_{\min} = 1$ ,  $n_{\min} = 2$ , and  $m_{\max} = n_{\max} = 20$ . For heterogeneous distributions we use  $g_m \propto m^{-\gamma_m}$  and  $p_n \propto n^{-\gamma_n}$ , with  $\gamma_m = \gamma_n = 3$ ,  $m_{\min} = n_{\min} = 2$ , and  $m_{\max} = n_{\max} = 100$ . The star markers indicate when  $\zeta = 1$  in the limit  $\epsilon \rightarrow 0$ , using Eq. (6.32). Irregularities of the level curves are due to the discrete nature of  $g_m$  and  $p_n$ , not to numerical errors. (d)-(f) Time evolution of the fraction of infected nodes for different strategies, with  $\epsilon = 10^{-2}$  and  $\nu \in \{1, 1.88, 3\}$ , corresponding to the three empty markers in (a). We use  $\lambda = 3\lambda_c$  in each case. The random strategy corresponds to  $s_m = 1 - \epsilon$  for all  $m$ , and  $f_{n,i}$  is a binomial distribution with probability  $\epsilon$ .

In the Supplementary Information, we show that

$$\lim_{\epsilon \rightarrow 0} \zeta = \frac{\beta(n', i')(n' - i')\langle n \rangle}{i' \langle \beta(n, 1)n(n-1) \rangle m_{\max}} + \frac{\langle m(m-1) \rangle}{\langle m \rangle m_{\max}}, \quad (6.31)$$

where  $(n', i')$  is the pair that maximizes the ratio  $R(n, i)$ , restricted to  $i > 0$ , in the limit  $\epsilon \rightarrow 0$ . For general  $\epsilon$ , we need to solve numerically the optimization problem as discussed in the previous sections.

Interestingly, with  $\beta(n, i) = \lambda i^\nu$ ,  $\zeta$  is independent of  $\lambda$ , since  $\Phi \propto \lambda$ . As a consequence,  $\zeta$  is agnostic to the underlying phase of the system (healthy, bistable, or endemic). Equation (6.31) simplifies to

$$\lim_{\epsilon \rightarrow 0} \zeta = \frac{i'^{\nu-1}(n' - i')\langle n \rangle}{\langle n(n-1) \rangle m_{\max}} + \frac{\langle m(m-1) \rangle}{\langle m \rangle m_{\max}}. \quad (6.32)$$

In Fig. 6.11, we illustrate how  $\zeta$  varies as we change  $\nu$ ,  $\epsilon$ , and the underlying structure. For homogeneous memberships and group sizes [Fig. 6.11(a)], we see that the influential groups strategy performs better as soon as the contagion process is sufficiently nonlinear ( $\nu \approx 2$ ); for highly nonlinear contagions ( $\nu \approx 4$ ), the influential group strategy is much more effective, with  $\zeta$  up to 100. When considering heterogeneous memberships, but still homogeneous group sizes [Fig. 6.11(b)], the influential spreaders strategy performs better for moderately nonlinear contagions ( $\nu \lesssim 2.8$ ), otherwise the influential groups

strategy is still a better choice. Finally, considering a heterogeneous  $p_n$  as well [Fig. 6.11(c)] helps the performance of the influential groups strategy, especially for larger  $\epsilon$ .

When picking a pair  $(\epsilon, \nu)$  far from  $\zeta = 1$  [dashed lines in Figs. 6.11(a), 6.11(b) and 6.11(c)], the strategy that maximizes  $\dot{I}$  invades the system faster, as can be seen in Figs 6.11(d) and 6.11(f). However, sufficiently close to  $\zeta = 1$ , maximizing  $\dot{I}$  does not necessarily imply that  $I(t)$  will be larger for all  $t > 0$ . For instance, in Fig. 6.11(e),  $\zeta \approx 1$ , but the influential spreader strategy is slightly better. Therefore, one must be careful when interpreting the results of Fig. 6.11. One way to improve on our approach would be to consider higher-order temporal derivatives of  $I$  to assess which strategy performs best, or refine the optimization procedure by trying to maximize these higher-order derivatives as well.

Interestingly, Figs. 6.11(d)–(f) suggest that the initial speed,  $\dot{I}(0)$ , roughly correlates with the time taken by the disease to infect a given fraction of the population, a metric that has been used to measure influence for SI and SIR dynamics [154, 174, 269].

Figure 6.11(f) also illustrates a particular feature of highly nonlinear contagions : the time to reach the stationary state can be excessively long for suboptimal strategies, despite  $\lambda = 3\lambda_c$ . In this regime, the initial conditions have a much more important impact on the capacity of the contagion to invade the system, especially considering the possibility of stochastic extinction in real systems due to finite size.

To understand why this is the case, we show in Fig. 6.12 the time evolution of  $f_{n_{\max}, i}(t)$ , the distribution for the number of infected in the largest groups, when the nodes are initially infected at random. For up to  $t = 50$ , only a few nodes are typically infected in the largest groups. Then we see the apparition of a bimodal distribution around  $t = 80$ , where either almost all nodes of the largest groups are infected or almost all nodes are susceptible. So instead of having a homogeneous transition, where all the largest groups have a similar state for all  $t$ , they diffuse asynchronously to states where almost all nodes are infected, at which point equilibrium is reached. This transition is driven by stochastic fluctuations, which explains the slow take-off. Conversely, the influential group strategy is by design optimizing the group configurations to facilitate this transition, which explains its good performance in Fig. 6.11(f).

It is also worth mentioning that the transition in Fig. 6.12 hinges on the nonlinearity of the contagion process—as soon as the number of infected nodes is sufficiently large in a group, nonlinear effects boost infections, resulting in almost all nodes being infected. Thus, nonlinear effects can be important *locally*, even though the *global* prevalence is still pretty low. With mean-field approaches that ignore dynamical correlations, nonlinear effects—or contributions from higher-order interactions—kick in only for a sufficiently large global prevalence [177].

These results again highlight the importance of considering an accurate description of the inner dynamics of groups when studying hypergraph contagions. In the context of influence maximization, optimizing group configurations is a crucial component ; one should not focus exclusively on identifying the most central nodes. Ultimately, an optimal strategy would capitalize on the synergy of these two important aspects.



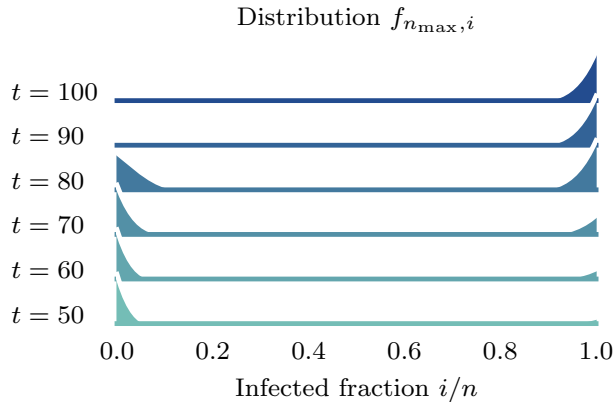


FIGURE 6.12 – Time evolution of the distribution  $f_{n_{\max},i}$  when using the random strategy. The structure, the dynamics and the parameters are identical to the ones of Fig. 6.11(f).

## 6.8 Discussion

We have introduced group-based AMEs to describe hypergraph contagions. Our framework is analytically tractable, allowing us to obtain closed form implicit expressions for the critical and tricritical points. In addition, we have shown that it describes the dynamical process with remarkable accuracy when compared with Monte Carlo simulations. Our formulation in terms of an infection rate function  $\beta(n, i)$  makes it extremely flexible, allowing us to consider arbitrary group distribution with large group interactions, contrarily to existing HMF theories [142, 147, 170] which instead require specifying the rule for each different type of interaction separately.

Motivated by simplicity and recent results [265], we analyzed in depth the consequences of a nonlinear infection rate function  $\beta(n, i) = \lambda i^\nu$ , highlighting the important role of influential groups in hypergraph contagions.

With our analytical results about the invasion and bistability thresholds, we were able to perform an exhaustive analysis of the phase transition and better understand the influence of a heterogeneous structure, both in terms of membership  $m$  and group size  $n$ . We found that the third moment of the membership distribution  $g_m$  plays a crucial role, with large  $\langle m^3 \rangle$  suppressing the onset of a discontinuous phase transition with a bistable regime, in line with other approaches [147, 170]. This is best exemplified for power-law membership distributions  $g_m \propto m^{-\gamma_m}$ , where  $\gamma_m = 3$  most suppresses bistability, and in the limit  $m_{\max} \rightarrow \infty$ , a discontinuous phase transition is only possible for  $\gamma_m > 4$ .

The phenomenon of mesoscopic localization, driven by the most influential groups, also has important consequences on the phase diagram, with the effects being enhanced by superlinear infection ( $\nu > 1$ ). In this case, the invasion threshold scales as  $\lambda_c \sim n_{\max}^{-\nu}$ , and for  $\lambda$  close to  $\lambda_c$ , infected nodes are found almost exclusively in the largest groups. This localization of the contagion thereby inhibits bistability by enforcing an endemic state with a very small global fraction of infected nodes.

Our approach, furthermore, provided novel insights concerning the problem of influence maximization for hypergraph contagions. We focused on the problem of maximizing the early spread, and proposed two strategies : allocating seeds to the influential spreaders (engineering  $s_m$ ), or to the influential groups (engineering  $f_{n,i}$ ). For various types of structures, the latter strategy performs better for contagions

that are sufficiently nonlinear, highlighting the key role of influential groups on the transient state of the system.

For the process we considered, the notion of influential groups to seed and sustain hypergraph contagions are mostly aligned—in both cases, the largest groups typically have a dominant role. In the case of influence maximization, however, we showed that a careful seed allocation is also essential to determine whether or not a group is influential. Moreover, a more realistic infection function  $\beta(n, i)$  that actually depends on  $n$  could affect which groups are most influential in both scenarios.

Our work constitutes a first step towards a better understanding of the role of higher-order interactions on the outreach of information spreading [244], and resonates with other recent theoretical findings on higher-order naming games, where big groups facilitate takeover of committed minorities in social convention [141]. Interestingly, AMEs thus provide an analytical avenue to study recent empirical results showing how social contagions and movements defy classic influence maximization [143]. As one example, networked counterpublics [144] are public spaces used by underrepresented groups to gather legitimacy and form tight-knit communities. Therein, nondominant forms of knowledge can still spread and reach widespread attention through dense communities (influential groups) despite the limited connectivity of their members (noninfluential spreaders). These results provide one more addition to the mounting evidence that groups of elementary elements are the foundational unit of many complex systems.

Many avenues are now left open to explore and broaden the applicability of our group-based AMEs. While we restrained ourselves to a particular nonlinear infection rate function  $\beta(n, i)$  and a constant recovery rate, other dynamical processes could be considered, each having their own phenomenology and a rich dynamical behavior. In the Supplementary Information for instance, we briefly discuss how our framework can be applied to threshold models of the form  $\beta(n, i) = \delta_{n-1, i}$ , but one could consider other traditional dynamical processes, such as voter models [259].

In the Supplementary Information, we provide a roadmap to include *structural* two-point correlations, but a thorough characterization of the impact of correlation patterns on bistability, mesoscopic localization and influence maximization is still lacking. The inclusion of *dynamical* correlation around nodes is a more tedious task that would require a fusion between degree-based [105, 106, 189] and group-based [125, 226, 266, 267] AMEs. This would allow to describe almost exactly *short-range* dynamical correlations, namely correlations between the states of nodes and their direct neighbors. Incorporating *long-range* correlations—beyond first neighbors—in AME frameworks, without a prohibitive computational time due to combinatorial explosion, is still an open problem.

Finally, many directions could be taken with regards to the influence maximization problem on hypergraphs. One avenue would be to analyze the notion of influential groups and influential spreaders from the perspective of centrality measures for hypergraphs [278]. Another would be to investigate the closely related problem of targeted immunization [131, 133, 231, 234].

## 6.9 Appendix : Contagion on real-world hypergraphs

### 6.9.1 Simulation of contagions

We used a standard Gillespie algorithm for the simulation of contagions on hypergraphs. We decompose the whole process into events  $j \in J$ , that each happens at rate  $\omega_j$ . The next event to happen is chosen with probability

$$P(j) = \frac{\omega_j}{\sum_{j \in J} \omega_j},$$

and the time step between two events  $\Delta t$  is distributed exponentially with mean  $\langle \Delta t \rangle = 1 / \sum_{j \in J} \omega_j$ .

There are two types of events : infection and recovery. On the one hand, all susceptible nodes in a group can be considered equivalent with regards to infection. Consequently, each group is chosen for an infection event with rate

$$\omega_{\text{inf}}(n, i) = (n - i)\beta(n, i).$$

Once a group is chosen for an infection event, one of the  $(n - i)$  susceptible nodes is chosen uniformly at random to become infected. On the other hand, all infected nodes perform a recovery event with rate  $\omega_{\text{rec}} = 1$ .

We store all possible events in an efficient data structure called a `SampleableSet` [260], where insertion, deletion and sampling of elements (events) all have a computational complexity  $O[\log \log (\omega_{\text{max}}/\omega_{\text{min}})]$  (see Appendix A), where  $\omega_{\text{max}}$  and  $\omega_{\text{min}}$  are respectively the maximal and minimal rate among  $\{\omega_j\}_{j \in J}$ . This makes the sampling and the updating of the data structure extremely fast, which is especially useful when  $\{\omega_j\}_{j \in J}$  spans multiple scales.

Once an event is performed—for instance a node recovers—we need to update the rate  $\omega_{\text{inf}}$  of all groups to which this node belongs. This is the most costly part of the algorithm, which unfortunately cannot be overcome. This essentially means the simulation procedure is slower for hypergraphs with large average excess membership  $\langle m(m - 1) \rangle / \langle m \rangle$ .

In Fig. 6.2 and 6.3, we compare the stationary state solutions from our formalism with estimates from Monte Carlo simulations. To compute estimates, we let the system relax during a burnin period  $\tau_b \in [10^2, 10^4]$  then we sample  $\mathcal{N} \in [10, 10^4]$  states, both depending on the size of the hypergraph and if multiple randomized hypergraphs are being used. Sampled states are separated by a decorrelation period  $\tau_d = 1$ .

To simulate contagions in the stationary state, we used two approaches, ordinary simulations and the quasistationary-state method.

**Ordinary simulation method.** With this approach, we simply let the simulation run and do not intervene. This is usually not the method of choice, especially for small hypergraphs near the invasion threshold, because finite size systems all eventually reach the absorbing state where all nodes are susceptible. This is, however, more practical to obtain the lower branch in Fig. 6.2(f), or faster for large hypergraphs, as in Fig. 6.3.

**Quasistationary-state method.** This approach aims at sampling the quasistationary distribution of the contagion process [68], which is defined as the probability distribution for all states in the limit

$t \rightarrow \infty$ , except for the absorbing state. We used the state-of-the-art method discussed in Ref. [68, 248], where we keep a history of past states (in our case up to 50 states). We update the history by removing one uniformly at random and storing the current state after each decorrelation period  $\tau_d \in [0.1, 1]$ . Each time the absorbing state is reached during the simulation, we pick a state from the history uniformly at random to replace the current one. This method is well suited for finite-size analysis, and especially useful for simulations on small hypergraphs, such as in Fig. 6.2.

### 6.9.2 Datasets

The simulations shown in Fig. 6.2 and Fig. 6.3 run on two different empirical social structures that encode different types of social higher-order interactions. Here, we briefly describe the nature of these two datasets and the techniques used to construct the associated higher-order structures.

**Face-to-face interactions in a French primary school.** Originally collected in Ref. [271] as part of the *SocioPatterns* [2] collaboration, this dataset contains information of face-to-face interactions between children of a primary school in Lyon recorded over 2 days. Participants are given wearable sensors (placed on their chests), and a contact is detected whenever two sensors are in close proximity (1.5m). The initial temporal resolution of this dataset is 20s, but contacts have been further pre-processed as in Ref. [142] in order to construct a static hypergraph from the temporal sequence of interactions. In particular, considering each child as a node, we aggregated different snapshots using a temporal window of 15 minutes and computed all the maximal cliques appearing in each window. Cliques were then aggregated across the entire time range, retaining only those that appeared at least 3 times, and finally “promoted” to groups. Some properties of the obtained structure are reported in the caption of Fig. 6.2.

**Co-authorship in computer science.** *DBLP* [1] is an online bibliography containing information on major computer science journals and proceedings. This dataset, already pre-processed in Ref. [25] (from the release 3, 2017), consists in a list of publications and respective authors that naturally calls for higher-order representations [235]. In particular, each author corresponds to a node and any collaboration of  $n$  (co-)authors in a single publication corresponds to a group of size  $n$ . We constructed a hypergraph by aggregating all the resulting groups together, but without considering single author publications (these have been removed in order to avoid disconnected nodes). In addition, the original dataset contained 1 831 127 nodes and 2 954 518 groups, which is too large to perform simulations on a personal computer in a reasonable time. Therefore, we obtained a sub-hypergraph by performing a breadth-first search, starting from a random group, then visiting all groups at a maximum distance of 2 when considering the one-mode projection of the original hypergraph on the groups. This ensures that the resulting sub-hypergraph is connected. Some properties of the obtained structure are reported in the caption of Fig. 6.3.

### 6.9.3 Randomization and data augmentation

In Fig. 6.2 and 6.3, we make use of randomized versions of the original hypergraph. In Fig. 6.2, we also use *expanded* versions, where the size of the network is increased by a factor  $x$ . In all cases, we use the same procedure ( $x = 1$  if the hypergraph is not expanded).

Let us first note  $\mathbf{m} = [m_1, m_2, \dots]$  and  $\mathbf{n} = [n_1, n_1, \dots]$  the membership sequence and the group size sequence of the original hypergraph, i.e., the list for the membership of each node and the list for the size of each group. From these sequences, we create two expanded sequences  $\mathbf{m}'$  and  $\mathbf{n}'$ , which are formed of  $x$  copies of  $\mathbf{m}$  and  $\mathbf{n}$  respectively. This can be seen as the membership and group size sequences for a hypergraph that is  $x$  times larger.

For each expanded sequence, we create a stub list. For instance, for the node  $j$  of the expanded hypergraph, we include  $m'_j$  copies of the label  $j$  in the stub list for the nodes. Similarly, we include  $n'_\ell$  copies of the label  $\ell$  in the stub list for the groups. By definition, these two stub lists are of the same length,  $M'$ , which corresponds to the number of *edges* in the bipartite representation of the hypergraph. We can thus shuffle them and match the entries of both lists, thereby assigning nodes to groups—or equivalently creating edges between nodes and groups in the bipartite representation of the hypergraph.

We then remove multi-edges (nodes assigned multiple times to a same group) by performing edge swaps [95]. We then perform  $M'$  additional edge-swap attempts at random—picking two random edges, swapping the groups, and accepting the swap if it does not create multi-edges. This ensures the uniformity of the generation process (see the Supplementary Information). The resulting hypergraph is a randomized version of the original hypergraph, expanded by a factor  $x$ .

## 6.10 Appendix : Influential groups solutions

An intuitive approach to solve the problem would be to sort all pairs  $(n, i)$  in decreasing order of their  $R(n, i)$  values (for  $i > 0$ ), then fill  $\tilde{f}_{n,i}$  up to 1 following this order, until  $I = \epsilon$ , or more directly until  $\eta$  reaches the value prescribed by  $\epsilon$ . However, this approach does not account for the fact that one may encounter multiple times a same  $n$  value before the condition  $I = \epsilon$  is reached. For instance, let us assume  $(n, i)$  is the next the pair with the highest value  $R(n, i)$ , but there exists a pair  $(n, i')$  with  $R(n, i') \geq R(n, i)$  and we have already assigned  $\tilde{f}_{n,i'} = 1$ . What is the best option ?

1. Discard the  $(n, i)$  pair.
2. Fill the associated  $\tilde{f}_{n,i}$  up to 1 and decrease the value of  $\tilde{f}_{n,i'}$  accordingly.

It turns out that an optimal solution is constructed by choosing one or the other depending on certain conditions. Option 1 is chosen whenever  $i < i'$ , because it can only reduce the total contribution to  $\Phi$ . If  $i > i'$ , we assign a new cost-effective ratio to the pair  $(n, i)$ , accounting for the fact that we need to decrease  $\tilde{f}_{n,i'}$  :

$$\hat{R}(n, i) = \frac{iR(n, i) - i'R(n, i')}{i - i'}.$$

This can be interpreted as the cost-effective ratio for the additional infected nodes  $(i - i')$  that we add to the configuration. Note that  $\hat{R}(n, i)$  can be negative, which is not a problem : this only means that infecting these nodes decreases the objective function  $\Phi$ . If  $\hat{R}(n, i)$  is still the highest ratio when compared with the ratios from other available pairs, option 2 is chosen.

The procedure to find an optimal set  $\tilde{\mathcal{F}}^\star$  is presented in algorithm 1, where we use a priority queue  $Q$  to keep sorted the pairs  $(n, i)$  in decreasing order of their cost-effective ratio. Using a binary heap for  $Q$ , inserting any pair  $(n, i)$  or removing the pair with maximum cost-effective ratio takes  $O(\log |Q|)$  elementary operations in the worst case, where  $|Q|$  is the number of elements in the queue. Since

$|Q| \leq n_{\max}^2$ , insertions and removals have a complexity  $O(\log n_{\max})$ . In the worst case, we will perform  $O(n_{\max}^2)$  insertions and removals before the condition  $I = \epsilon$  is reached, hence the worst-case complexity for algorithm 1 is  $O(n_{\max}^2 \log n_{\max})$ .

We also need to precompute  $R(n, i)$  for all  $n, i > 0$  and get  $\mathcal{F}^\star$  from  $\tilde{\mathcal{F}}^\star$ , both  $O(n_{\max}^3)$ . Moreover, solving for  $\eta$  is  $O(m_{\max})$ , so the overall computational complexity of the influential groups optimization procedure is  $O(m_{\max} + n_{\max}^3)$ .

---

**Algorithm 1** Optimization for influential groups

---

**Input :**  $p_n, R(n, i), \eta$

**Output :**  $\tilde{\mathcal{F}}^\star$

```

1:  $\tilde{f}_{n,i} \leftarrow 1$  if  $i = 0$ , else 0,  $\forall n, i$ 
2:  $\psi \leftarrow \eta\langle n \rangle$  ▷ Initial resources
3:  $i_{\text{opt}} \leftarrow \emptyset$  ▷ Associative array
4:  $Q \leftarrow \emptyset$  ▷ Priority queue
5: for all  $(n, i)$  s.t  $n \leq n_{\max}, p_n > 0$  and  $1 \leq i \leq n$  do
6:    $R \leftarrow R(n, i)$ 
7:    $Q \leftarrow (n, i)$  with priority  $R$ 
8: end for
9: while  $Q$  is not empty and  $\psi > 0$  do ▷ Highest priority element
10:   $(n, i) \leftarrow Q$ 
11:  if  $n \notin i_{\text{opt}}$  then
12:     $\tilde{f}_{n,i} \leftarrow \min[1, \psi/(ip_n)]$ 
13:     $\tilde{f}_{n,0} \leftarrow \tilde{f}_{n,0} - \tilde{f}_{n,i}$ 
14:     $\psi \leftarrow \psi - \tilde{f}_{n,i}ip_n$ 
15:     $i_{\text{opt}}[n] \leftarrow i$ 
16:  else if  $i > i_{\text{opt}}[n]$  then
17:     $i' = i_{\text{opt}}[n]$ 
18:     $\hat{R} \leftarrow [iR(n, i) - i'R(n, i')]/(i - i')$ 
19:    if  $\hat{R} > R$  for all  $R \in Q$  then
20:       $\tilde{f}_{n,i} \leftarrow \min\{1, \psi/[p_n(i - i')]\}$ 
21:       $\tilde{f}_{n,i'} \leftarrow \tilde{f}_{n,i'} - \tilde{f}_{n,i}$ 
22:       $\psi \leftarrow \psi - \tilde{f}_{n,i}p_n(i - i')$ 
23:       $i_{\text{opt}}[n] \leftarrow i$ 
24:    else
25:       $Q \leftarrow (n, i)$  with priority  $\hat{R}$ 
26:    end if
27:  end if
28: end while

```

---

## 6.11 Supplementary Information

### 6.11.1 AME correction for correlations

It is straightforward to adapt our formalism to include two-point structural correlations between node memberships and group sizes. Let  $P(m, n)$  be the joint probability distribution for the membership of a node and the size of a group at the endpoint of an edge in the bipartite representation of the hypergraph. Marginalization results in

$$\sum_m P(m, n) = \frac{np_n}{\langle n \rangle}$$

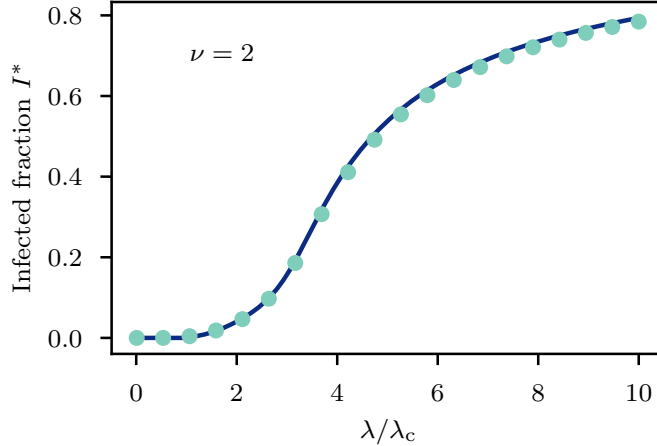


FIGURE 6.13 – Assessing dynamical correlation around nodes by removing nodes with a large membership. We performed the same experiment as in Fig. 6.3(c), but we used a homogeneous membership distribution  $g_m \approx 0.25\delta_{m,3} + 0.75\delta_{m,4}$  with the same mean  $\langle m \rangle \approx 3.75$ . The group size distribution is unchanged.

and

$$\sum_n P(m, n) = \frac{mg_m}{\langle m \rangle}.$$

We thus define  $P(m|n)$ , the conditional probability for a node to be of membership  $m$  in a group of size  $n$ , and  $P(n|m)$ , the conditional probability for a group to be of size  $n$  taking a random group to which a node of membership  $m$  belongs,

$$P(m|n) \equiv \frac{P(m, n)\langle n \rangle}{np_n} \quad \text{and} \quad P(n|m) \equiv \frac{P(m, n)\langle m \rangle}{mg_m}. \quad (6.33)$$

From these two definitions, it is sufficient to replace  $r(t) \mapsto r_m(t)$  and  $\rho(t) \mapsto \rho_n(t)$  into Eq. (6.4) to account for two-point structural correlations, where

$$r_m(t) = \frac{\sum_{n,i} \beta(n, i)n^{-1}(n-i)f_{n,i}P(n|m)}{\sum_{n,i} n^{-1}(n-i)f_{n,i}P(n|m)}, \quad (6.34a)$$

$$\rho_n(t) = \frac{\sum_m (m-1)r_m s_m P(m|n)}{\sum_m s_m P(m|n)}. \quad (6.34b)$$

While this correction leaves unaltered the dimension of the system of ordinary differential equations, the computational complexity associated to the numerical integration of Eq. (6.4) increases from  $O(n_{\max}^2 + m_{\max})$  to  $O(n_{\max}^2 m_{\max})$  per time step.

For a more in-depth study of the effect of structural correlations on hypergraph contagion models, see Ref. [170]. To include dynamical correlations around nodes, it would necessitate a much more thorough reformulation of our formalism, for instance by combining the partitioning in groups with the node partitioning typically used in degree-based AMEs [105, 106, 178, 189].

### 6.11.2 Further assessment of the effect of dynamical correlation around nodes

We conjecture that the disparity between the AMEs and the simulations in Fig. 6.3(c) is caused by the presence of nodes with very large membership. The dynamical correlations around these nodes become important at high prevalence and it is not correctly captured by our AMEs.

To support our claim, we reproduced the simulation and the results of the AMEs in Fig. 6.13, keeping the same group size distribution, but using a homogeneous membership distribution. The mean membership  $\langle m \rangle \approx 3.75$  is preserved.

The phase diagram is quite different without the heterogeneous membership distribution—the phase transition is smeared due to mesoscopic localization near the invasion threshold. However, the point is that the agreement is better at high prevalence, which suggests that the nodes with large membership had a role to play in the disparity.

### 6.11.3 Jacobian of the mean-field map

All quantities are assumed to have reached the stationary state. As shown in Sec. 6.6, the stationary state solutions are completely characterized by the mean field  $r$ . Therefore, one just needs to obtain solutions  $r$  that respect

$$r = \mathcal{M}[\rho(r)] .$$

A useful quantity to find the fixed points numerically and analyze their stability is the Jacobian

$$J_r = \frac{d\mathcal{M}}{dr} = \frac{d\mathcal{M}}{d\rho} \frac{d\rho}{dr} .$$

For the first term, we define the following quantities

$$\begin{aligned} u_1(\rho) &= \sum_{n,i} \beta(n,i)(n-i)f_{n,i}(\rho)p_n , \\ v_1(\rho) &= \sum_{n,i} (n-i)f_{n,i}(\rho)p_n , \\ \frac{du_1}{d\rho} &= \sum_{n,i} \beta(n,i)(n-i) \frac{df_{n,i}}{d\rho} p_n , \\ \frac{dv_1}{d\rho} &= \sum_{n,i} (n-i) \frac{df_{n,i}}{d\rho} p_n , \end{aligned}$$

and write

$$\frac{d\mathcal{M}}{d\rho} = \frac{1}{v_1} \frac{du_1}{d\rho} - \frac{u_1}{v_1^2} \frac{dv_1}{d\rho} .$$

After some manipulations, one can write

$$\frac{df_{n,i}}{d\rho} = -f_{n,i} \sum_{j>0} f_{n,j} \Delta_{n,j} + f_{n,i} \Delta_{n,i} ,$$

where

$$\Delta_{n,i}(\rho) = \sum_{j=0}^{i-1} \frac{1}{\beta(n,j) + \rho} .$$



For the second term we follow the same steps, starting with the definitions

$$\begin{aligned} u_2(r) &= \sum_m m(m-1)s_m g_m, \\ v_2(r) &= \sum_m m s_m g_m, \\ \frac{du_2}{dr} &= \sum_m m(m-1) \frac{ds_m}{dr} g_m, \\ \frac{dv_2}{dr} &= \sum_m m \frac{ds_m}{dr} g_m, \end{aligned}$$

such that we can write

$$\frac{d\rho}{dr} = \frac{u_2}{v_2} + r \left[ \frac{1}{v_2} \frac{du_2}{dr} - \frac{u_2}{v_2^2} \frac{dv_2}{dr} \right].$$

We insert the following in the previous definitions

$$\frac{ds_m}{dr} = -\frac{m}{(1+mr)^2}.$$

Combining all equations, we are able to compute the Jacobian  $J_r$  in  $O(m_{\max} + n_{\max}^2)$  elementary operations.

#### 6.11.4 Threshold models

All quantities are assumed to have reached the stationary state. For the phase transition analysis we performed in Sec. 6.6, we assumed that the infection rate  $\beta(n, i) > 0 \forall i > 0$  to simplify the expressions. Therefore, the results we obtained for the invasion and bistability threshold break down if we use a threshold model, for instance of the form  $\beta(n, i) \equiv \lambda \delta_{n-1, i}$ , as used in Ref. [147].

To understand why our equations for the critical and tricritical points are not valid, let us consider the two level threshold model

$$\beta(n, i) = H(i-1)\varepsilon + (\lambda - \varepsilon)\delta_{n-1, i}.$$

If  $\varepsilon \rightarrow 0$ , we recover  $\beta(n, i) \equiv \lambda \delta_{n-1, i}$ . Using Eq. (6.15), we have for small  $\varepsilon$

$$\lambda_c \approx \frac{\langle m \rangle}{\langle m(m-1) \rangle} \varepsilon^{2-n}.$$

Therefore, for  $n > 2$  and finite mean excess membership, the invasion threshold diverges in the limit  $\varepsilon \rightarrow 0$ , as can be seen in Fig. 6.14. The contagion is then a pure critical mass model [75], i.e., a vanishing initial fraction of infected nodes cannot lead to an endemic state and there is no invasion threshold.

The only way to have a nondiverging invasion threshold when  $\varepsilon \rightarrow 0$  is to have a diverging excess membership, typically when  $g_m \sim m^{-\gamma_m}$ . A complete analysis of the critical properties of this regime was performed in Ref. [147] using a heterogeneous mean-field approximation, showing that both continuous and discontinuous phase transitions are possible. The results they obtained are in agreement with our framework. Indeed, if we introduce  $\beta(n, i) \equiv \lambda \delta_{n-1, i}$  in Eq. (6.8), we have

$$\begin{aligned} f_{n, i} &= f_{n, 0} \frac{n!}{(n-i)! i!} \rho^{i-1} (\rho + \lambda \delta_{n, i}), \\ &= f_{n, 0} \left\{ \frac{n!}{(n-i)! i!} \rho^i + \lambda \delta_{n, i} \rho^{n-1} \right\}, \end{aligned}$$

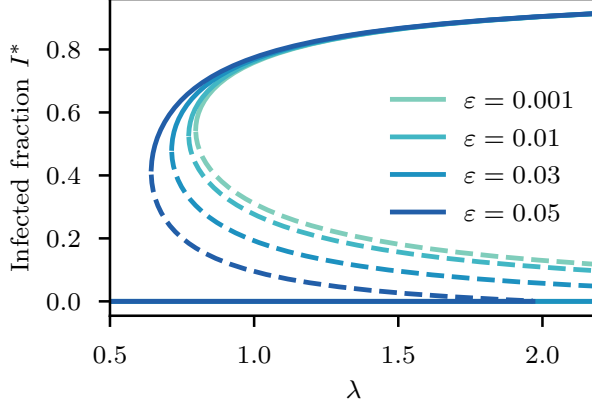


FIGURE 6.14 – Stationary fraction of infected nodes for a threshold model. Solid lines and dashed lines represent stable and unstable solutions respectively. The infection rate function is  $\beta(n, i) = H(i - 1)\varepsilon + (\lambda - \varepsilon)\delta_{n-1, i}$ . The structure is fixed with  $g_m = \delta_{m,6}$  and  $p_n = \delta_{n,3}$ .

with

$$f_{n,0} = [(1 + \rho)^n + \lambda\rho^{n-1}]^{-1}.$$

If we further assume that  $p_n \equiv \delta_{n,\bar{d}}$  (only one type of interaction), then

$$r = \frac{\lambda\bar{d}\rho^{\bar{d}}}{(1 + \rho)^{\bar{d}} - 1} \approx \lambda\rho^{\bar{d}-1},$$

for  $\rho \ll 1$ . This is the same functional form as used in Ref. [147] to describe their mean-field quantities in the stationary state. Our approach would therefore predict the same critical exponents.

### 6.11.5 Uniformly sampling stub-labeled simple bipartite graphs

The rationale behind the additional  $M'$  edge-swap attempts in Sec. 6.9.3 is to ensure the uniformity of the generation process. Indeed, we want to sample uniformly from the ensemble of all *stub-labeled* [95] simple bipartite graphs with fixed membership and group size sequences. By performing a stub matching and eliminating the multi-edges, the generating process is not uniform.

However, by using arguments similar to the ones presented in Ref. [95], it is straightforward to show that performing random edge swaps as described in the main text creates an ergodic Markov chain that has a uniform stationary distribution on the ensemble of graphs of interest. Thus, if a sufficient number of edge swaps attempts are performed, the Markov chain approximates the stationary distribution and the resulting hypergraph is properly randomized.

To assess whether the chain is sufficiently close to the stationary distribution, it is standard to look at a number of scalar measures and to make sure the chain has converged [99]. In Fig. 6.15, we show such results for the bipartite clustering coefficient  $C_b$ , as defined by Robins and Alexander [243]. We see that for both datasets, whether we start the chain with the original hypergraph or with a random configuration from the stub-matching process, the distribution eventually settles around a stationary value. Moreover, we see that  $\sim M'$  edge-swap attempts ( $\sim 1$  Markov-chain step per edge) is usually sufficient to reach the stationary distribution, and that using a random configuration from the stub-matching process puts us closer to the stationary distribution initially.

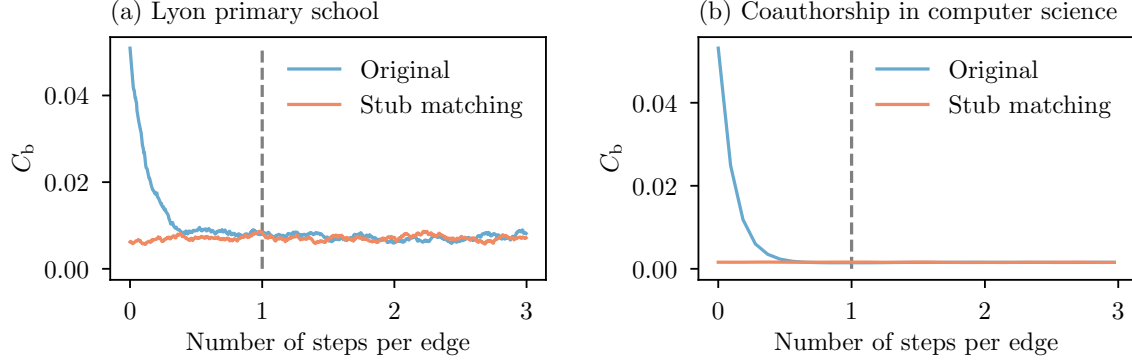


FIGURE 6.15 – Characterization of the convergence for the randomization procedure. We define a Markov chain where at each step we perform an edge-swap attempt on the bipartite representation of the hypergraph. If this does not create multi-edges, the step is accepted, otherwise we keep the previous configuration. The stationary distribution of this Markov chain is uniform on all *stub-labeled* simple bipartite graphs. To characterize the mixing time, we look for the convergence of the bipartite clustering coefficient  $C_b$ , as defined by Robins and Alexander [243]. We initialize the Markov chain with the original hypergraph constructed from the data (blue lines) and a random configuration produced by stub matching (red lines). (a) Hypergraph constructed from high-resolution face-to-face contact data from a primary school in Lyon (see Sec. 6.9.2). The bipartite graph representation contains 2853 edges and we perform a measure of the clustering coefficient after each 10 edge-swap attempts. (b) Hypergraph constructed from coauthorship data in computer science (see Sec. 6.9.2). The bipartite graph representation contains 215 434 edges and we perform a measure of the clustering coefficient after each  $2 \times 10^4$  edge-swap attempts.

### 6.11.6 Influence maximization : limit case

All quantities are evaluated at  $t = 0$ . An important limit case for influence maximization is when the number of nodes to infect  $I = \epsilon$  is very small. Fortunately, we can evaluate  $\zeta$  in the limit  $\epsilon \rightarrow 0$ . Let us develop  $\Phi$  to first order in  $\epsilon$  for the optimal solution of both strategies.

For the influential spreader strategy,

$$s_m \simeq 1 - \frac{\epsilon \delta_{m, m_{\max}}}{g_{m_{\max}}},$$

and

$$q \simeq \frac{\epsilon m_{\max}}{\langle m \rangle}.$$

Therefore, to first order in  $\epsilon$ ,

$$\Phi_{\mathcal{S}}^{\star} \propto \frac{\epsilon \langle \beta(n, 1) n(n-1) \rangle m_{\max}}{\langle m \rangle}. \quad (6.35)$$

For the influential group strategy,

$$s_m \simeq 1 - \eta m \Rightarrow \eta \simeq \frac{\epsilon}{\langle m \rangle} \quad \text{and} \quad u \simeq \epsilon \frac{\langle m(m-1) \rangle}{\langle m \rangle}.$$

Also,  $\tilde{f}_{n,i} = 0$  for all  $n$  and  $i > 0$ , except for a particular pair  $(n', i')$  that maximizes the ratio  $R(n, i)$ . Equation (6.28) implies

$$\tilde{f}_{n', i'} \simeq \frac{\epsilon \langle n \rangle}{\langle m \rangle i' p_{n'}},$$

and therefore to first order in  $\epsilon$

$$\Phi_{\mathcal{F}}^{\star} \propto \epsilon \frac{\langle n \rangle}{\langle m \rangle} R(n', i') + \epsilon \langle \beta(n, 1) n(n-1) \rangle \frac{\langle m(m-1) \rangle}{\langle m \rangle^2}, \quad (6.36)$$

where more formally

$$R(n', i') = \max(\{R(n, i) | 0 < i \leq n \wedge 2 \leq n \leq n_{\max}\}) .$$

Finally, combining Eqs. (6.35) and (6.36) we obtain Eq. (6.31) in the main text.

## Conclusion et perspectives

Les réseaux d'ordre supérieur sont sans aucun doute *le* sujet à la mode en ce moment en science des réseaux [22]. Avant de participer à la mouvance, j'ai longtemps réfléchi à la pertinence d'adopter cette perspective pour mes travaux, centrés sur les processus stochastiques de contagion. Il n'y a pas si longtemps de cela, la structure mésoscopique des systèmes complexes était la plupart du temps caractérisée à partir d'un graphe et de sa structure *communautaire*, un sujet brièvement abordé à la section 1.1.1. La terminologie était différente — on parlait alors de *réseaux modulaires* — mais les processus dynamiques possédaient généralement des propriétés similaires [73, 104, 215, 226, 293]. Est-ce que l'étude des dynamiques sur réseaux d'ordre supérieur est seulement un effet de mode ? Ou est-ce réellement une perspective riche pour la science des réseaux, et plus généralement, pour la science des systèmes complexes ? Je suis du second avis.

Précédemment, les groupes de noeuds avaient un rôle de second plan, l'accent étant mis sur les conséquences d'une structure modulaire, comme l'augmentation du coefficient d'agrégation. Les réseaux d'ordre supérieur permettent d'inclure plus directement et activement les groupes dans la dynamique. Le groupe a une signification qui va au-delà d'une plus grande fréquence des interactions, elle définit un ensemble d'éléments en relation, et cette relation peut se manifester de différentes façons [277]. Elle peut servir à représenter des corrélations spatiotemporelles entre les individus, comme la coprésence dans un endroit, ou indiquer une plus grande pression sociale exercée par un groupe versus celle associée aux dyades. C'est pourquoi il faut utiliser une représentation qui élève naturellement le groupe au même pied d'égalité que les noeuds. L'utilisation du langage des hypergraphes ou des complexes simpliciaux n'est pas nécessaire — nous-même dans cette thèse nous favorisons la représentation bipartite des hypergraphes —, mais l'important est de considérer les groupes comme une unité fondamentale à part entière des systèmes complexes.

Dans cette thèse, nous avons démontré l'importance de porter une attention particulière aux interactions de groupe pour les processus de contagion. Aux chapitres 3 et 4, nous avons étudié des processus de contagion simple — linéaire — sur réseaux d'ordre supérieurs. Même si le processus résultant est isomorphe à une dynamique sur graphes formés de cliques, la perspective centrée sur les groupes nous a menés à introduire et caractériser un nouveau phénomène de localisation mésoscopique, pour lequel un sous-ensemble des groupes participe de manière active à l'épidémie. L'élément clé ayant permis cette analyse est la description précise des corrélations dynamiques à l'intérieur des groupes. De plus, nous avons montré au chapitre 4 qu'une intervention ciblée sur les groupes peut avoir des répercussions très différentes de ce que l'on pourrait s'attendre en utilisant la perspective des graphes, notamment à cause de la localisation mésoscopique.

Au chapitre 5, nous avons étudié une dynamique de contagion complexe combinant les concepts de

seuil (dose infectieuse minimale) et de temps d'exposition hétérogène dans un environnement. Notre modèle se décrit naturellement sous la forme d'un réseau biparti, avec les individus — les noeuds — participant à différents environnements — les groupes. Notre résultat principal montre qu'une dynamique complexe associée à la coprésence des individus peut mener à une forme universelle de contagion non linéaire, dont le taux d'infection varie en loi de puissance avec le nombre d'individus infectés dans l'environnement. Cela met en relation deux types de contagions complexes discutés à la section 2.3, les contagions basées sur la notion de seuil, comme le modèle de Granovetter, et celles basées sur un taux d'infection (ou force d'infection) non linéaire.

Au chapitre 6, nous avons poussé plus loin l'analyse des contagions non linéaires sur réseaux d'ordre supérieur grâce aux équations maîtresses approximées. Encore une fois, nous avons montré qu'une description précise des corrélations dynamiques à l'intérieur d'un groupe est essentielle pour bien prédire les propriétés du système à l'équilibre. Les effets de localisation mésoscopique deviennent encore plus importants pour les contagions non linéaires et affectent grandement le portrait de phase. Mais probablement le résultat le plus intéressant est celui associé à la maximisation de l'influence : nous montrons que pour une contagion suffisamment non linéaire, bien choisir la configuration initiale des groupes influents est plus important pour envahir rapidement une population que de cibler les noeuds influents. Il est alors essentiel de traiter les groupes comme des objets irréductibles, ce qu'amène naturellement la représentation en réseaux d'ordre supérieur.

## Vers une caractérisation complète des contagions sur réseaux d'ordre supérieur

Le contenu de cette thèse fut limité aux projets complétés jusqu'à maintenant. Toutefois, nous avons identifié au fil des chapitres plusieurs problèmes ouverts pour lesquels nous avons entamé des travaux préliminaires. Faisons un bref tour d'horizon.

Aux chapitres 3 et 4, nous introduisons le concept de localisation *mésoscopique* sur les groupes qui se distingue d'une localisation *microscopique* centrée sur les noeuds uniquement. Toutefois, l'échelle mésoscopique est très large selon notre définition : elle touche tout ce qui est entre l'échelle microscopique (un noeud et ses voisins) et l'échelle macroscopique (l'ensemble du système). En ce sens, les travaux de cette thèse ne couvrent que le premier sous-niveau de l'échelle mésoscopique, soit les groupes de taille finie (bien que certains groupes puissent contenir un grand nombre de noeuds). Plus généralement, pour bien comprendre la localisation sur les regroupements de noeuds, il faudra explorer les différents sous-niveaux de l'échelle mésoscopique. C'est une observation bien établie que les systèmes complexes possèdent une structure avec plusieurs niveaux d'organisation [124, 290] ou couches [5, 123], ce qui pourrait inspirer une caractérisation *hiérarchique* du phénomène de localisation.

Au chapitre 5, nous obtenons un modèle *effectif*<sup>2</sup> de contagion non linéaire en moyennant le temps d'exposition dans un environnement. Cela nous amène à nous poser la question suivante : *est-ce que les contagions non linéaires peuvent décrire d'autres systèmes pour lesquels les hétérogénéités locales ont été moyennées ?* Il se trouve que c'est le cas. Par exemple, en prenant bien soin de suivre adéquatement les corrélations dynamiques à l'intérieur des groupes, une contagion linéaire dont le taux d'infection à l'intérieur de chaque groupe  $\beta$  est une variable aléatoire est équivalente (sous certaines conditions) à une contagion non linéaire.

---

2. Au sens où l'*effet* résultant est équivalent à une contagion non linéaire, même si les mécanismes sous-jacents ne sont pas directement un taux (ou une probabilité) d'infection non linéaire.

Finalement, au chapitre 6, nous ouvrons la porte à deux généralisations théoriques de nos travaux. (1) La localisation mésoscopique fut caractérisée de manière numérique seulement pour les contagions non linéaires. Nous travaillons en ce moment sur une adaptation de l'analyse du chapitre 3 à ce cas plus général. (2) Les corrélations dynamiques autour des noeuds peuvent être importantes, comme le suggère la figure 6.3. Afin d'avoir une description précise des contagions — et plus généralement des processus stochastiques discrets — sur réseaux d'ordre supérieur, il faut combiner les équations maîtresses approximées centrées sur les noeuds et celles centrées sur les groupes pour former des équations maîtresses approximées généralisées (*generalized approximate master equations* ; GAMEs). Nous avons réussi à formuler de telles équations sans qu'elles soient prohibitives d'un point de vue computationnel, il ne reste plus qu'à trouver la bonne approche pour les présenter à la communauté scientifique.

### Les contagions complexes comme modèles effectifs en épidémiologie

Tout au long de cette thèse, nous nous sommes concentrés sur l'analyse théorique des processus de contagion. Lorsque l'on considère des problèmes plus appliqués, par exemple en épidémiologie, il faut faire face à plusieurs difficultés, dont l'absence d'une connaissance complète de la structure des interactions et les données disponibles sont souvent bruitées ou erronées. C'est pourquoi l'épidémiologie sur réseaux est rarement utilisée pour faire des prédictions en temps réel sur l'évolution d'une épidémie<sup>3</sup>. L'exception est probablement les modèles de métapopulation [14, 60, 157, 231, 290]. La structure du réseau correspond plutôt au flot de mobilité des individus entre chaque sous-population, mais dans ces modèles, les sous-populations sont généralement homogènes et suivent la loi d'action de masse.

Comment peut-on importer les connaissances et les nouvelles intuitions apportées par l'épidémiologie sur réseaux dans des modèles suffisamment simples pour être déployées ? Le chapitre 5 de cette thèse, nos travaux en cours et certaines études récentes [126] suggèrent que les contagions non linéaires pourraient être de bons modèles effectifs, faisant abstraction de l'hétérogénéité et des détails microscopiques grâce à un taux d'infection plus général. On pourrait alors utiliser certaines formes de contagion non linéaires pour incorporer de manière implicite la complexité des systèmes sociaux dans les modèles épidémiologiques.

Cette avenue génère pour l'instant beaucoup plus de questions que de réponses. Jusqu'à quel point peut-on moyenniser l'hétérogénéité des contacts, le risque associé à différents environnements, etc., et toujours avoir un modèle qui reste fidèle au processus initial ? Quelle forme de contagion non linéaire doit-on utiliser ? Comment inférer de manière robuste les paramètres d'un modèle de contagion non linéaire en contexte réaliste ? En répondant à ces questions, on peut espérer une meilleure compréhension des risques et un meilleur contrôle sur les épidémies. Bref, des questions difficiles, et un espoir fou, les bases idéales pour un programme de recherche stimulant.

---

3. Ce qui n'empêche pas l'épidémiologie sur réseaux d'être grandement utile pour étudier l'influence d'une structure sociale complexe et réaliste sur la propagation [181].

# Annexes



## Annexe A

# Simulation des processus de contagion sur réseaux

Article original :

**Efficient sampling of spreading processes on complex networks using a composition and rejection algorithm**

**Guillaume St-Onge**, Jean-Gabriel Young, Laurent Hébert-Dufresne et Louis J Dubé

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

Référence : *Comput. Phys. Commun.* **240**, 30 (2019)

© 2019 Elsevier (§A.3-A.11)<sup>1</sup>

---

1. Ces sections contiennent le contenu original de l'article. Celui-ci n'a été modifié que pour se conformer au format exigé par la Faculté des études supérieures et postdoctorales de l'Université Laval.

*Any one who considers arithmetical methods of producing random digits is, of course, in a state of sin.*

—John von Neumann [285]

## A.1 Avant-propos

Bien que les approches théoriques permettent de mieux comprendre le système, il est toujours sage de valider les résultats obtenus par des simulations numériques du processus stochastique que l'on tente de modéliser. Les quantités que l'on cherche à valider sont souvent des espérances sur les états  $\mathbf{x}$  du système, de la forme  $\langle f(\mathbf{x}) \rangle$ . On peut donc simuler plusieurs trajectoires du processus stochastique pour obtenir un échantillon  $\{\mathbf{x}^{(j)}\}_{j=1}^{\mathcal{N}}$ , où  $\mathcal{N}$  est le nombre de trajectoires. On peut alors estimer l'espérance comme suit :

$$\langle f(\mathbf{x}) \rangle \approx \frac{1}{\mathcal{N}} \sum_{j=1}^{\mathcal{N}} f(\mathbf{x}^{(j)}) .$$

Dans ce chapitre, nous nous concentrons sur la simulation des contagions simples sur réseaux d'interaction par paires (graphes). Par contre, le concept central de notre approche — l'échantillonnage par composition et réjection — fut réutilisé dans différents contextes, pour simuler des contagions complexes sur réseaux d'ordre supérieur (voir section 6.9), mais aussi dans le cadre de l'inférence temporelle des réseaux [42, 302]. En effet, la méthode proposée résout un problème beaucoup plus général, soit celui d'échantillonner de manière efficace une distribution hétérogène et *dynamique* (qui change à étape du processus stochastique).

Notre solution est encapsulée dans une structure de données que l'on nomme un `SamplableSet` [260]. Cette structure est d'ailleurs utilisée par un nouvel outil pour la simulation des processus de contagion, `epipack` [188].

Symbole	Description
$N$	Nombre de noeuds du réseau
$M$	Nombre de liens du réseau
$S, I, R$	États possibles pour chaque noeud du réseau (susceptible, infecté, retiré)
$\mathbf{x}$	Vecteur des états des noeuds, où $x_i \in \{S, I, R\}$ est l'état du noeud $i$
$\lambda$	Taux de transmission de la maladie aux voisins
$\mu$	Taux de récupération pour les noeuds infectés
$u_\alpha$	Processus microscopique indépendant (transmission de la maladie $i \rightarrow j$ ou récupération pour un noeud $i$ )
$w_\alpha$	Taux associé au processus $u_\alpha$
$\Omega_T(\mathbf{x})$	Ensemble des processus de transmission pour un état $\mathbf{x}$
$\Omega_R(\mathbf{x})$	Ensemble des processus de récupération pour un état $\mathbf{x}$
$\Omega(\mathbf{x})$	Ensemble des processus microscopiques indépendants pour un état $\mathbf{x}$ ( $\Omega_T \cup \Omega_R$ )
$W(\mathbf{x})$	Taux global, soit la somme des taux de tous les processus indépendants
$\tau(\mathbf{x})$	Temps durant lequel le système reste dans l'état $\mathbf{x}$
$k_i$	Degré du noeud $i$
$k_I$	Degré d'un noeud nouvellement infecté

$\omega_i(\mathbf{x})$	Taux associé à l'ensemble des processus impliquant le noeud $i$
$\omega_{\min}$	Taux minimal associé à un noeud
$\omega_{\max}$	Taux maximal associé à un noeud
$g_\beta$	Ensemble de noeuds regroupés selon des taux $\omega_i$ similaires
$\mathcal{P}$	Partition des noeuds $\{g_\beta\}_{\beta=1}^q$
$W_\beta(\mathbf{x})$	Taux associé à l'ensemble des noeuds $g_\beta$

TABLEAU A.1 – Glossaire des symboles importants utilisés dans l'annexe A.

## A.2 Résumé

Avoir en main des algorithmes de simulations efficaces est d'une importance capitale pour l'étude des processus stochastiques de contagion sur réseaux complexes. En nous appuyant sur des résultats analytiques provenant de la science des réseaux, nous discutons de l'effet de la structure sur l'efficacité de certains algorithmes couramment utilisés. Malgré une idée répandue, soit que ces algorithmes requièrent  $O(\log N)$  ou même  $O(1)$  opérations élémentaires pour chaque étape du processus — où  $N$  est le nombre de noeuds —, nous montrons que le nombre d'opérations croît de manière polynomiale avec  $N$  pour les réseaux denses et les réseaux creux hétérogènes. Cela affecte significativement le temps de calcul pour les simulations sur des réseaux de grande taille. Pour résoudre ce problème, nous proposons une méthode regroupant les actions des noeuds du réseau, combinée avec un algorithme de composition et de réjection. Cette procédure permet d'échantillonner la prochaine étape du processus stochastique en utilisant en moyenne  $O[\log \log N]$  opérations élémentaires pour un réseau arbitraire. Nous introduisons notre méthode dans le cadre des processus de Markov, mais il est aussi possible de l'adapter à un certain nombre de processus non markoviens. Nos résultats ont le potentiel de faciliter considérablement l'étude d'un large éventail de dynamiques sur réseaux.

## A.3 Abstract

Efficient stochastic simulation algorithms are of paramount importance to the study of spreading phenomena on complex networks. Using insights and analytical results from network science, we discuss how the structure of contacts affects the efficiency of current algorithms. We show that algorithms believed to require  $O(\log N)$  or even  $O(1)$  operations per update—where  $N$  is the number of nodes—display instead a polynomial scaling for networks that are either dense or sparse and heterogeneous. This significantly affects the required computation time for simulations on large networks. To circumvent the issue, we propose a node-based method combined with a composition and rejection algorithm, a sampling scheme that has an average-case complexity of  $O[\log(\log N)]$  per update for general networks. This systematic approach is first set-up for Markovian dynamics, but can also be adapted to a number of non-Markovian processes and can enhance considerably the study of a wide range of dynamics on networks.

## A.4 Introduction

Stochastic processes in a discrete state space are useful models for many natural and human-related complex systems. Considering that complex and heterogeneous connectivity patterns form the backbone of these systems, the study of dynamical processes on networks has grown in popularity in the last decades. Spreading processes are a prime example [163, 231, 286]. They are used to model a wide range of phenomena related to propagation among a population. While disease epidemics [12] is the most commonly studied process, other important examples include social contagions [176, 210] and even beneficial epidemics [26].

Many analytical approaches have been developed to study the outcome of spreading processes on complex networks, using mean field [19, 37, 209, 279], moment closure [78, 193, 253, 268], percolation mapping [159, 217, 228] and message passing techniques [149, 254] to name a few (see Refs. [163, 231, 286] for recent overviews). However, most results hold only for random networks and tree-like structures, or stand as approximations for general networks. This undeniably contributes to our understanding of real systems, but any conclusions drawn from these approaches need to be supported by exact, robust, numerical simulations.

However, an important limitation of numerical simulations is certainly their computational cost, which generally increases with the size of the system. Spreading processes are a class of nonequilibrium models that exhibit a phase transition in the thermodynamic limit (infinite size systems). This particular feature has motivated the need for simulation algorithms that can handle network with very large number of nodes  $N$ . This is especially true for the study of anomalous phenomena, such as the localization of the epidemic [45, 109, 180, 230], Griffith phases and smeared phase transitions [62, 64, 195, 213, 223, 268], typically observed in large-size networks. Therefore, the development of stochastic simulation algorithms that can efficiently tackle large networks of varied structures is unescapable, although it offers a considerable, albeit fascinating, computational challenge. It is also critical to the foundations of the field of spreading processes, one of the cornerstones of network science.

One standard formulation of spreading on networks is in terms of a continuous-time stochastic process. A widely used numerical approach involves a temporal discretization using a finite time step. Despite its simplicity, this is both computationally inefficient—requiring  $O(N)$  operations per time step—and prone to discretization errors [86]. Instead, the state-of-the-art methods are based on the Doob-Gillespie’s algorithm [102], which produces statistically exact sequences of states. Many studies have been dedicated to the improvement of its efficiency [101, 111, 258, 297], and from these ideas have emerged some implementations for spreading processes [61, 89, 163, 192, 281]. However, these approaches can still be inefficient for certain structures, for instance dense or sparse and heterogeneous networks.

In this work, we propose an efficient method for the simulation of Markovian spreading processes that builds on the Doob-Gillespie’s algorithm. It hinges on a composition and rejection scheme [258] that requires only  $O[\log(\log N)]$  operations per update for general networks. We demonstrate that it provides a formidable improvement compared to other algorithms whose computation time can scale polynomially with the network size. An implementation of the algorithm is available online [261].

## A.5 Markovian spreading processes on networks

For the sake of simplicity, we consider simple graphs of  $N$  nodes and  $M$  edges, although the methods discussed can all be applied to directed, weighted, and multi-layer networks. We consider spreading processes defined as continuous Markov processes with discrete states  $\mathbf{x}$ . We can focus on canonical epidemiological models to understand the methods—other compartmental models can also be accommodated, for instance by adding accessible states for the nodes, but numerical approaches are similar. Let us therefore denote the state of each node  $i$  as  $x_i \in \{S, I, R\}$ , susceptible, infected or recovered respectively. If infected node  $i$  and susceptible node  $j$  are connected, node  $i$  transmits the disease to node  $j$  at rate  $\lambda$ . Node  $i$  spontaneously recovers at rate  $\mu$ , and becomes recovered if it develops immunity against the spreading disease, or susceptible otherwise. From this framework, we distinguish two models with different properties in the limit  $t \rightarrow \infty$ .

**SIS model** Infected nodes become directly susceptible after recovery (they do not develop immunity). For finite size networks, the absorbing state with all nodes susceptible, where the system remains trapped, is eventually visited. However, for infinite size networks<sup>2</sup>, there exists some threshold  $\lambda_c$  for fixed  $\mu$  such that the system reaches an endemic state, where a stationary density of the nodes remains infected on average, for all  $\lambda > \lambda_c$  [see Fig. A.1(a)]. This stationary density of infected nodes is hereafter referred as the *prevalence*.

---

2. In practice, this is observed for large (finite size) networks.

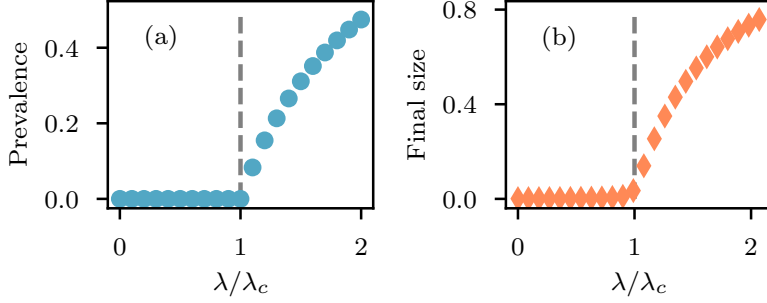


FIGURE A.1 – Typical phase transition of the order parameters for the two canonical models of spreading. The recovery rate is fixed to  $\mu = 1$  while the transmission rate  $\lambda$  is varied. Simulations were performed using the composition and rejection algorithm (see Sec. A.6.2) on a single realisation of the  $G(N, M)$  random graph ensemble [81], with  $N = 10^5$  and  $M = 5 \times 10^5$ . (a) Average prevalence in the endemic state for the SIS model. (b) Average final size for an outbreak of the SIR model, starting with an initial infected node density of  $10^{-3}$ . In both cases, the error bars associated to the standard deviation are hidden by the markers.

**SIR model** Recovered nodes develop immunity. After a certain time, all nodes are either susceptible or recovered. Similarly to the SIS model, there exists a threshold value  $\lambda_c$  above which even an infinitesimal initial density of infected nodes will ultimately lead to a macroscopic number of recovered nodes that scales with network size [see Fig. A.1(b)]. The final density of recovered nodes is hereafter referred as the *final size*.

Threshold estimates are provided in A.9 for uncorrelated random networks with an arbitrary degree distribution.

## A.6 Exact simulation algorithms

Spreading processes can be decomposed into a set  $\Omega(\mathbf{x}) = \{u_\alpha\}$  of independent processes for each state  $\mathbf{x}$  of the system. A process  $u_\alpha$  can be the transmission of a disease from node  $i$  to  $j$  or the recovery of an infected node. We can distinguish the set of transmission processes  $\Omega_T(\mathbf{x})$  and the set of recovery processes  $\Omega_R(\mathbf{x})$  such that  $\Omega_T \cap \Omega_R = \emptyset$  and  $\Omega_T \cup \Omega_R = \Omega$ . To establish a coherent framework for the stochastic simulation algorithms, we consider that the execution of a process is an event that *may* change the state  $\mathbf{x}$  of the system, but it is not required to do so. For instance, the transmission of a disease from node  $i$  to  $j$  is a process that implies  $x_j \mapsto I$  if  $x_j = S$ . However, if  $x_j \in \{I, R\}$  prior to the transmission, then the state of the system stays the same. Considering or not the transmission of a disease between two infected nodes as an ongoing process is only an algorithmic decision; from our definition, an infected node transmits the disease to all of its neighbors irrespectively of their states.

Since every possible transmission and recovery for a state  $\mathbf{x}$  is an independent process  $u_\alpha$  with rate  $w_\alpha \in \{\lambda, \mu\}$ , the total rate of events is defined as

$$W(\mathbf{x}) = \sum_{u_\alpha \in \Omega(\mathbf{x})} w_\alpha. \quad (\text{A.1})$$

The inter-event time  $\tau(\mathbf{x})$  before the execution of any process is exponentially distributed

$$P(\tau|\mathbf{x}) = \exp(-W(\mathbf{x})\tau)dt, \quad (\text{A.2})$$

and the executed process is chosen proportionally to its *propensity*, i.e. its rate  $w_\alpha$

$$P(u_\alpha|\mathbf{x}) = \frac{w_\alpha}{W(\mathbf{x})}. \quad (\text{A.3})$$

In this section, we introduce and discuss different numerical methods to sample sequences of states based on the previous equations. To compare the methods and give insights on their behavior for large system, we will provide expressions for the theoretical number of operations required on average to perform a state transition, i.e.  $\mathbf{x} \mapsto \mathbf{x}'$  with  $\mathbf{x}' \neq \mathbf{x}$ . The expressions are *upper bounds* to the number of operations required on average and are expressed using the big  $O$  notation.

### A.6.1 Doob-Gillespie's algorithm (direct method)

The direct method consists of the determination of the next process  $u_\alpha$  and the time at which it will be executed. It can be summarized by these steps :

1. Determine the inter-event time  $\tau(\mathbf{x})$  using Eq. (A.2).
2. Determine the executed process  $u_\alpha$  using Eq. (A.3).
3. Update the state  $\mathbf{x} \mapsto \mathbf{x}'$  according to the chosen process  $u_\alpha$  (if  $\mathbf{x}' \neq \mathbf{x}$ ), and update the time  $t \mapsto t + \tau(\mathbf{x})$ .
4. Update, create, or remove processes resulting from the execution of process  $u_\alpha$ .
5. Update  $W(\mathbf{x})$  and  $P(u_\alpha|\mathbf{x})$ , then return to step 1.

There have been various implementations of this direct procedure that have been discussed for spreading processes on static networks [61, 86, 163, 192].

For general networks, the number of ongoing processes is  $O(M+N)$ . Efficient implementations typically use a binary-tree data structure to store the processes and keep updated  $P(u_\alpha|\mathbf{x})$  [101, 192]. This allows the insertion and the retrieval of processes in  $O(\log N)$  operations. The costly processes involve the infection of a degree  $k_I$  node. To perform this update, the required number of operations is

$$O(k_I \log N) \tag{A.4}$$

and is associated with the storage of future transmission processes.

A fact often overlooked is that the degree  $k_I$  of a newly infected node can be large on average. For dense networks, it scales as  $k_I \sim N$ ; for sparse and heterogeneous networks near the phase transition,  $k_I \sim N^\xi$  on average with  $0 \leq \xi \leq 1$  (see A.9). This is problematic, since phase transitions are central to many studies.

### A.6.2 Node-based method

Another type of implementation has been suggested [61, 89]. To sample among the infection processes, one selects a node  $i$  among the infected nodes, proportionally to its degree  $k_i$ , then selects one of its neighbors  $j \in \partial i$  randomly and infects it if  $x_j = S$ . The sampling of recovery processes is simply done by selecting an infected node uniformly. Using this scheme, one does not have to look through all the neighbors of an infected node to check for new processes  $u_\alpha$ , associated with the  $O(k_I)$  operations in the direct method.

We introduce what we call the *node-based* method in the spirit of this scheme. The idea is to regroup the propensity of all processes associated with a node into a single propensity in order to prevent their enumeration. The probability of selecting a process  $u_\alpha$  performed by node  $i$  is then factorized as  $P(u_\alpha, i|\mathbf{x}) = P(u_\alpha|i, \mathbf{x})P(i|\mathbf{x})$ . The probability of selecting a node is written

$$P(i|\mathbf{x}) = \frac{\omega_i}{W(\mathbf{x})}, \tag{A.5}$$

where

$$\omega_i(\mathbf{x}) = \delta(x_i, I)(\lambda k_i + \mu) \tag{A.6}$$

is the total propensity for node  $i$  to be involved in a transmission or recovery event,  $\delta(a, b)$  is the Kronecker delta and  $k_i$  is the degree of node  $i$ . The probability of selecting a process  $u_\alpha$  performed by

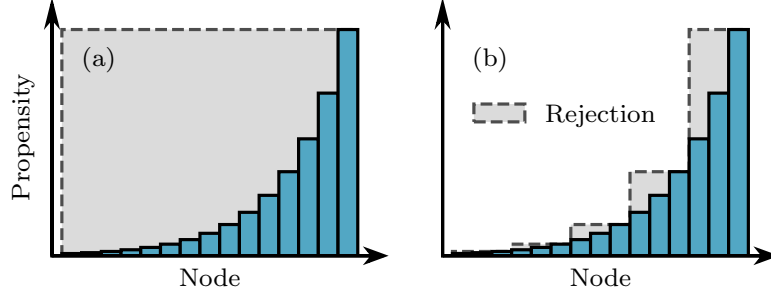


FIGURE A.2 – Rejection area (gray) compared to the acceptance area (blue), for the selection of a node with the node-based method. (a) Rejection sampling. (b) Composition and rejection sampling.

node  $i$  is formally written

$$P(u_\alpha|i, \mathbf{x}) = \frac{w_\alpha}{\omega_i(\mathbf{x})} . \quad (\text{A.7})$$

In practice, we use the probability that the chosen process is a transmission

$$P(u_\alpha \in \Omega_T|i, \mathbf{x}) = \frac{\lambda k_i}{\lambda k_i + \mu} , \quad (\text{A.8})$$

to select the type of process. If it is a transmission, we select a random neighbor and infect it ; if it is a recovery, the node becomes susceptible or recovered depending on the spreading model used. Note that the total rate of events [Eq. (A.1)] has an equivalent definition in terms of the propensities of the nodes, namely

$$W(\mathbf{x}) = \sum_{i=1}^n \omega_i(\mathbf{x}) . \quad (\text{A.9})$$

An important property of this scheme in the context of spreading processes is that Eq. (A.6) is only dependent on the state of the node  $i$ . Hence, when the state of a node  $i$  changes due to the execution of a process, we do not need to update the propensity  $\omega_j$  of neighboring nodes  $j \in \partial i$ , but only  $\omega_i$ . We discuss two implementations for this method.

### Rejection sampling

A simple approach to select a node  $i$  according to  $P(i|\mathbf{x})$  is to use rejection sampling. First, one needs to determine the maximal propensity possible  $\omega_{\max}$  for a node, in this case being

$$\omega_{\max} = \lambda k_{\max} + \mu , \quad (\text{A.10})$$

where  $k_{\max}$  is the maximal degree of a node in the network. Second, one needs to keep updating an array of pairs of elements  $\mathcal{A}$ , the pairs being node labels and their propensity  $(i, \omega_i)$ —we keep only pairs with nonzero propensities  $\omega_i$ . Finally, one draws a pair  $(i, \omega_i)$  at random from  $\mathcal{A}$  and accepts it with probability

$$P[(i, \omega_i)|\omega_{\max}] = \frac{\omega_i}{\omega_{\max}} . \quad (\text{A.11})$$

The complete procedure for the execution of a process using the node-based method with rejection sampling is presented in Algorithm 2. Although the SIS model was considered, one only needs to change step 19 by  $S \mapsto R$  for it to be compliant with the SIR model.

It is straightforward to verify that most steps are  $O(1)$  in Algorithm 2, except possibly the rejection sampling portion (steps 3 to 9). The average acceptance probability for a state  $\mathbf{x}$  is  $\langle \omega_i|I; \mathbf{x} \rangle / \omega_{\max}$ , where

$$\langle \omega_i|I; \mathbf{x} \rangle = \lambda \langle k|I; \mathbf{x} \rangle + \mu , \quad (\text{A.12})$$

---

**Algorithm 2** Node-based method with rejection sampling for an update of the SIS model

---

**Input :**  $\mathbf{x}, \mathcal{A}, W$

**Output :**  $\tau$  and updated input

```

1: choose  $r_1 \in [0, 1)$  uniformly at random
2:  $\tau \leftarrow -\frac{\ln r_1}{W}$ 
3: select  $(i, \omega_i) \in \mathcal{A}$  uniformly at random
4: choose  $r_2 \in [0, 1)$  uniformly at random
5: if  $r_2 < \frac{\omega_i}{\omega_{\max}}$  then
6:   choose node  $i$ 
7: else
8:   go back to step 3
9: end if
10: choose  $r_3 \in [0, 1)$  uniformly at random
11: if  $r_3 < \frac{\lambda k_i}{\lambda k_i + \mu}$  then ▷ transmission process
12:   choose  $j \in \partial i$  uniformly at random
13:   if  $x_j = S$  then
14:      $x_j \leftarrow I$ 
15:      $\mathcal{A} \leftarrow \mathcal{A} \cup \{(j, \omega_j)\}$ 
16:      $W \leftarrow W + \omega_j$ 
17:   end if
18: else ▷ recovery process
19:    $x_i \leftarrow S$ 
20:    $\mathcal{A} \leftarrow \mathcal{A} \setminus \{(i, \omega_i)\}$ 
21:    $W \leftarrow W - \omega_i$ 
22: end if

```

---

is the average propensity for infected nodes in state  $\mathbf{x}$  and  $\langle k|I; \mathbf{x} \rangle$  stands for the average degree of infected nodes in state  $\mathbf{x}$ . Since the number of required operations follows a geometric distribution, the average number of operations for an update from a state  $\mathbf{x}$  is

$$O(\omega_{\max} / \langle \omega_i | I; \mathbf{x} \rangle). \quad (\text{A.13})$$

For networks with a homogeneous degree distribution, we have  $\langle \omega_i | I; \mathbf{x} \rangle \sim \omega_{\max}$  for all states, leading to a small number of operations. However, this rejection sampling scheme is vulnerable to cases where the propensity ranges over multiple scales—more specifically, when  $\langle \omega_i | I; \mathbf{x} \rangle \ll \omega_{\max}$  for typical states  $\mathbf{x}$ . This happens when  $k_{\max}$  is large, which is common for heterogeneous degree distribution, such as  $P(k) \sim k^{-\gamma}$  (see A.9). To circumvent this, we must update and sample efficiently from a heterogeneous distribution of propensities  $P(i|\mathbf{x})$  [Eq. (A.5)].

### Composition and rejection for multiscale propensity

The high rejection probability of the rejection sampling for heterogeneous propensities is illustrated by the gray portion in Fig. A.2(a). The problem is the uniform proposal distribution on  $\mathcal{A}$  (step 3 of Algorithm 2), which is a poor choice for a heterogenous distribution  $P(i|\mathbf{x})$ . To improve upon rejection sampling, we need an algorithm that *systematically* constructs a proposal distribution that upper bounds the rejection probability, as illustrated in Fig. A.2(b).

We propose to use a method of *composition and rejection*, similarly to Ref. [258] for biochemical reaction networks, inspired by Ref. [70]. It is also a direct improvement over the 2-group method proposed in Ref. [61]. The idea is to create a partition of nodes  $\mathcal{P} = \{g_\beta\}$  with similar propensities  $\omega > 0$ , with  $\beta \in \{1, 2, \dots, q\}$ . Once a node  $i$  gets infected (or more generally gets a propensity  $\omega_i > 0$ ), it is assigned to a group  $g_\beta$ —in our implementation, the pair  $(i, \omega_i)$  is stored in an array  $\mathcal{A}_\beta$ . The probability to select a node  $i$  is then factorized as  $P(i|\mathbf{x}) = P(i|g_\beta, \mathbf{x})P(g_\beta|\mathbf{x})$ .



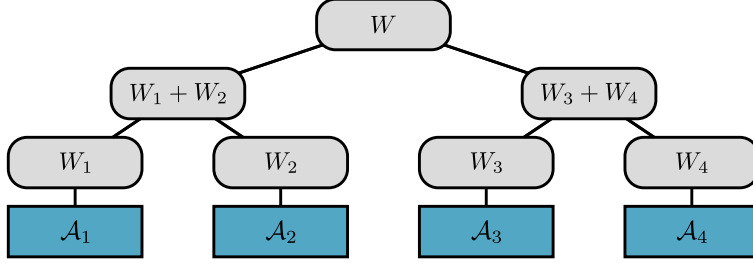


FIGURE A.3 – Decision tree  $\mathcal{T}$  used to select the arrays  $\{\mathcal{A}_\beta\}$ .

The probability of selecting a group  $g_\beta$  is

$$P(g_\beta|\mathbf{x}) = \frac{W_\beta(\mathbf{x})}{W(\mathbf{x})}, \quad (\text{A.14})$$

where  $W_\beta(\mathbf{x}) \equiv \sum_{i \in g_\beta} \omega_i(\mathbf{x})$  is the propensity associated to the group of nodes. In practice, to select an array  $\mathcal{A}_\beta$  proportionally to  $W_\beta$ , we implement a binary decision tree  $\mathcal{T}$  as illustrated in Fig. A.3, where each leaf points to an array  $\mathcal{A}_\beta$ . Starting from the root, it takes  $O(\log q)$  operations to choose one of the leaves. Once a process  $u_\alpha$  is chosen and executed, we update the array  $\mathcal{A}_\beta$ , the propensity  $W_\beta$  and recursively the parent values in the tree—one notes that the root value is in fact  $W(\mathbf{x})$ . Again,  $O(\log q)$  operations are needed for this task.

The probability of selecting a node within a group is

$$P(i|g_\beta, \mathbf{x}) = \frac{\omega_i(\mathbf{x})}{W_\beta(\mathbf{x})}. \quad (\text{A.15})$$

If the partition  $\mathcal{P}$  is wisely chosen, nodes are selected efficiently using rejection sampling, replacing  $\omega_{\max}$  in Eq. (A.11) by a group specific maximal propensity  $\omega_{\beta\max}$ .

A systematic approach to construct the partition  $\mathcal{P}$  is to impose a minimum acceptance probability, say one half. This leads to an average number of operations upper-bounded by 2. Let us define the minimal propensity as

$$\omega_{\min} = \lambda k_{\min} + \mu, \quad (\text{A.16})$$

where  $k_{\min}$  is the minimal degree in the network. We impose that the  $\beta$ -th group allows nodes with propensity  $\omega \in [2^{\beta-1}\omega_{\min}, 2^\beta\omega_{\min})$ , except for the last group allowing  $\omega \in [2^{q-1}\omega_{\min}, \omega_{\max}]$ . The group specific maximal propensity is thus

$$\omega_{\beta\max} = \min(2^\beta\omega_{\min}, \omega_{\max}), \quad (\text{A.17})$$

and the number of group required is

$$q = \max\left(1, \left\lceil \log_2 \left( \frac{\omega_{\max}}{\omega_{\min}} \right) \right\rceil\right). \quad (\text{A.18})$$

The complete procedure for the execution of a process using the node-based method with composition and rejection sampling is presented in Algorithm 3.

For general networks with  $\lambda/\mu < \infty$ , the ratio of extreme propensities  $\omega_{\max}/\omega_{\min} = O(N)$ . Therefore, the average-case complexity per update is

$$O[\log(\log N)]. \quad (\text{A.19})$$

For networks with a maximal degree  $k_{\max}$  independent of  $N$ , the complexity is  $O(1)$ .

---

**Algorithm 3** Node-based method with composition and rejection sampling for an update of the SIS model

---

**Input :**  $\mathbf{x}, \mathcal{T}$

**Output :**  $\tau$  and updated input

```

1: get  $W$  from the root of  $\mathcal{T}$ 
2: choose  $r_1 \in [0, 1)$  uniformly at random
3:  $\tau \leftarrow -\frac{\ln r_1}{W}$ 
4: choose  $r_2 \in [0, 1)$  uniformly at random
5: select  $\mathcal{A}_\beta$  from  $\mathcal{T}$  using  $r_2$ 
6: select  $(i, \omega_i) \in \mathcal{A}_\beta$  uniformly at random
7: choose  $r_3 \in [0, 1)$  uniformly at random
8: if  $r_3 < \frac{\omega_i}{\omega_{\beta\max}}$  then
9:   choose node  $i$ 
10: else
11:   go back to step 6
12: end if
13: choose  $r_4 \in [0, 1)$  uniformly at random
14: if  $r_4 < \frac{\lambda k_i}{\lambda k_i + \mu}$  then ▷ transmission process
15:   choose  $j \in \partial i$  uniformly at random
16:   if  $x_j = S$  then
17:      $x_j \leftarrow I$ 
18:      $\mathcal{A}_{\beta'} \leftarrow \mathcal{A}_{\beta'} \cup \{(j, \omega_j)\}$  ▷  $j \in g_{\beta'}$ 
19:     update  $\mathcal{T}$ 
20:   end if
21: else ▷ recovery process
22:    $x_i \leftarrow S$ 
23:    $\mathcal{A}_\beta \leftarrow \mathcal{A}_\beta \setminus \{(i, \omega_i)\}$ 
24:   update  $\mathcal{T}$ 
25: end if

```

---

It is worth mentioning that our choice to impose a minimum acceptance probability of one half is not unique : one could impose an acceptance probability of  $a^{-1}$  with  $a > 1$  and obtain the same average-case complexity. For  $a \rightarrow 1$ , it increases the acceptance probability, but it also increases the number of groups required  $q$ . Therefore, there is a trade-off to consider when one tries to minimize the required number of operations. We made several trials with  $a \neq 2$ , but it has never resulted in noticeable improvements for the computation time.

### A.6.3 Event-driven method

Another type of approach for the simulation of spreading processes has been considered lately [89, 163]. The philosophy is similar to the next reaction method [101], originally proposed for the simulation of chemical reaction networks to improve upon the original Doob-Gillespie's algorithm. The principal concept of this scheme is to draw an inter-event time  $\tau_\alpha$  for each of the specific process  $u_\alpha$ , and execute the latter at time  $t_\alpha + \tau_\alpha$ , where  $t_\alpha$  is the absolute time when the process  $u_\alpha$  was created. Therefore, one focus on the execution time of each process independently, instead of inferring the global inter-event time  $\tau(\mathbf{x})$  and the first process to be executed among  $\Omega(\mathbf{x})$ , as in standard Gillespie method.

For Markovian dynamics, the inter-event time before the execution of a process  $u_\alpha$  is exponentially distributed

$$P(\tau_\alpha) = \exp(-w_\alpha \tau_\alpha) dt_\alpha . \tag{A.20}$$

However, it is important to stress that this approach can also be applied to non-Markovian spreading processes [89, 163].

To store and retrieve the processes efficiently, one can use a priority queue, where the highest priority element corresponds to the process  $u_\alpha$  with the lowest absolute execution time  $t_\alpha + \tau_\alpha$ . Recovery processes are eventually executed and can be stored directly in the priority queue. Transmission processes are stored if the inter-event time for the transmission is smaller than the inter-event time for the recovery of the infected node. Depending on the implementation, one can also verify *a priori* if a neighbor node will already be infected or recovered, and prevent the storage of the transmission process.

To compare this approach with our node-based schemes, we used the implementation provided by Ref. [163, 202], called the *event-driven method*, for which a detailed pseudocode is available for both the SIS and the SIR models. The interface is modular and general enough to be used for any networks, a requirement for our benchmark calculations. Let us mention that there are alternative implementations available, for instance the efficient implementation of Ref. [132] for the SIR model (see Supplemental Material A.11 for a comparison with our algorithm).

As for the standard Gillespie algorithm, a costly process for the event-driven method involves the infection of a degree  $k_I$  node, for which

$$O(k_I \log N) , \tag{A.21}$$

operations are required. While this quantity is probably not the most representative for the average computation time associated with the algorithm [see Fig. A.4], we can clearly identify the different impact of the term  $k_I$  for dense or sparse and heterogeneous networks.

## A.7 Efficiency of the stochastic simulation algorithms

We have described different schemes for the simulation of spreading processes, with different associated complexity for the update of the state  $\mathbf{x} \mapsto \mathbf{x}'$ . In this section, we show and discuss the impact on the computation time using synthetic networks. In Fig. A.4 we compile the results of our benchmark calculations, using the SIS model in the stationary state and complete sequences of the SIR model, starting with a small infected node density. We fix the recovery rate  $\mu = 1$  without loss of generality. Furthermore, we scale the transmission rate parameter  $\lambda = \eta\lambda_c$  according to the underlying threshold  $\lambda_c$  of the dynamics. The intent is to always have a similar prevalence (SIS) or final size (SIR) in our simulations as we tune the network structure : as we discuss in A.10, different values for the order parameter can affect the expected number of required operations.

Implementations for the two node-based methods—hereafter referred as to *rejection* and *composition and rejection*—are in C++, while the implementation of the event-driven method is in Python [163, 202]. This discrepancy of programming languages causes a multiplicative factor overhead of roughly 100 for the computation times of the event-driven method. To provide a fairer comparison, and since we are mostly interested in the scaling of the algorithms with the size of the network, we have rescaled the computation times obtained with the event-driven method to match the first marker of the composition and rejection method in each panel of Fig. A.4. For instance, if  $T_{ED}(N)$  [ $T_{CR}(N)$ ] represents the the computation time of the event-driven (composition and rejection) method for different sizes  $N$ , we perform the transformation  $T_{ED}(N) \mapsto T_{CR}(N_{\min})T_{ED}(N)/T_{ED}(N_{\min})$ , with  $N_{\min} = 10^3$  in Fig. A.4. A comparison without rescaling is also available in the Supplemental Material A.11.

### A.7.1 Computation time for homogeneous networks

As a model of homogeneous networks, we used the  $G(N, M)$  random graphs ensemble [81]. In the limit  $N \rightarrow \infty$ , the degree distribution is a binomial with  $\langle k \rangle = 2M/N$  and the degree of nodes are well represented by the mean value  $\langle k \rangle$ .

We observe in Fig. A.4(a) that the average computation time for the event-driven method roughly scales linearly with the average degree, in agreement with Eq. (A.21). For the SIR model in Fig. A.4(d), the dependence is less important, except for very large average degree.

As a comparison, both node-based methods are completely independent of the average degree [Fig. A.4(a) and A.4(d)]. This is in line with the two steps selection procedure for infection processes. A node-based

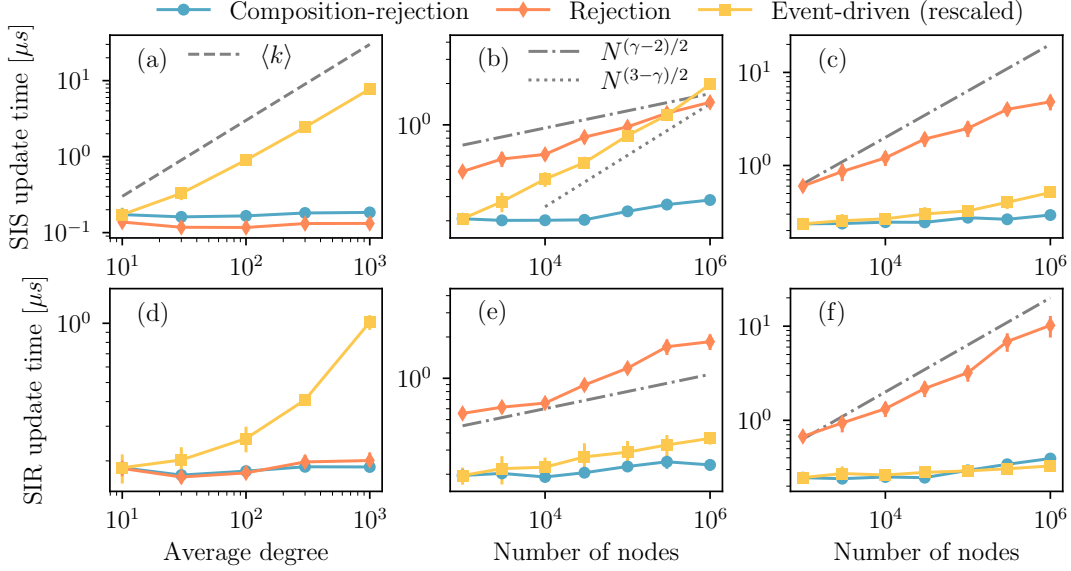


FIGURE A.4 – Average computation time for a single update of the state for spreading processes using different algorithms. Each marker is averaged over 10 simulations on 10 realizations of a random graph ensemble. The computation time of the event-driven method was rescaled to match the first marker of the composition and rejection method. (Upper row) Average over sequences of  $10^6$  state transitions for the SIS model in the stationary state. The systems had been thermalized using  $10^6$  transitions beforehand. (Lower row) Average over  $10^6/N$  complete sequences of the SIR model, starting with an initial infected node density of  $10^{-2}$ . (a) and (d)  $G(N, M)$  random graphs with fixed number of nodes  $N = 10^4$  and different average degree  $\langle k \rangle$ . The recovery and transmission rates used are  $\mu = 1$  and  $\lambda = 1.05\lambda_c$ . (b) and (e) Random graphs with an expected degree sequence [57, 203]. We used a power law expected degree distribution  $P(\kappa) \propto \kappa^{-\gamma}$  with  $\gamma = 2.25$ ,  $\kappa_{\min} = 3$  and  $\kappa_{\max} < \langle k \rangle N^{1/2}$ . The recovery and transmission rates used are  $\mu = 1$  and  $\lambda = 3\lambda_c$ . (c) and (f) Same as (b) and (e), but with  $\gamma = 3$ .

scheme should therefore be privileged whenever one wants to sample networks with large average degree, such as dense networks where  $\langle k \rangle = O(N)$ .

One also observes that the rejection method is slightly more efficient [Fig. A.4(a)] than the composition and rejection method : this is due to the simpler implementation of the rejection sampling, without the need for a composition step. As discussed in Sec. A.6.2, the average number of operations  $O(\omega_{\max}/\langle \omega_i | I; \mathbf{x} \rangle)$  is only problematic when propensities span multiple scales ; for homogeneous networks, the ratio  $\omega_{\max}/\langle \omega_i | I; \mathbf{x} \rangle$  is expected to be  $O(1)$  for all states.

### A.7.2 Computation time for heterogeneous networks

As a model of heterogeneous networks, we used random graphs with an expected degree sequence  $\{\kappa_i\}$  [57, 203], also called Chung-Lu graphs, where  $\kappa_i$  is the expected degree of node  $i$ . We used sequences drawn from a power-law distribution  $P(\kappa) \propto \kappa^{-\gamma}$  to generate heterogeneous networks, with  $\gamma = 2.25$  in Fig. A.4(b) and Fig. A.4(e), and  $\gamma = 3$  in Fig. A.4(c) and Fig. A.4(f).

We observe in Fig. A.4(b) that the computation time for the event-driven method scales polynomially with the number of nodes. In Fig. A.4(c), the computation time slightly increases, but with a much smaller exponent. This is explained by Eq. (A.21), with

$$k_I \sim N^{(3-\gamma)/2} \quad (\text{A.22})$$

near the phase transition (see A.9). For the SIR model in Fig. A.4(e) and A.4(f), the computation times are less influenced by the size of the network. In this case, the number of required operations

predicted in Eq. (A.21) by the most costly processes overestimates the average computation time (as also noted for homogeneous networks).

We observe that the rejection method scales polynomially with the number of nodes as well, but this time the scaling exponent is larger for moderately heterogeneous networks [Fig. A.4(c) and A.4(f)] than for very heterogeneous networks [Fig. A.4(b) and A.4(e)]<sup>3</sup>. This is roughly explained by Eq. (A.13), with  $\langle \omega_i | I; \mathbf{x} \rangle \sim \langle k^2 \rangle$  (see A.9) and  $\omega_{\max} \sim k_{\max}$ , leading to

$$\frac{\omega_{\max}}{\langle \omega_i | I; \mathbf{x} \rangle} \sim N^{(\gamma-2)/2}. \quad (\text{A.23})$$

Finally, we see that the computation time for the composition and rejection method is, for all practical purposes, independent of the number of nodes.

## A.8 Concluding remarks

We have introduced a stochastic simulation algorithm for the simulation of spreading processes on networks, combining a node-based perspective with the efficiency of composition and rejection sampling [261]. This algorithm requires  $O[\log(\log N)]$  operations per update, providing a superior scaling (or at worst, equivalent) to the other state-of-the-art methods. It is particularly well suited for the sampling of large and heterogeneous networks, since its average computation time is, for all practical purposes, independent of the network size or the density of edges.

For the SIR model specifically, our benchmark calculations reveal that the event-driven method scales efficiently as well, except for very dense networks. If computation time is a crucial issue, we suggest to compare the speed of our implementation [261] with highly optimized implementations of the event-driven method, such as the one provided by Ref. [132], or the recent hybrid approach proposed by Ref. [116], and choose the one that performs better for the particular case study.

Note that there is a complete branch of literature whose concern is the efficient sampling of the quasistationary distribution of states for processes with an absorbing state, such as the SIS model (see Refs. [61, 68, 186, 248] for instance). Indeed, finite size analysis of the critical phenomenon requires the sampling of sequences that do not fall on the absorbing state. In this paper, we have focused on the stochastic simulation algorithms that provide statistically exact sequences of states, which is also fundamental to sample the quasistationary distribution. Therefore, our work contributes indirectly to this line of study.

Despite the fact that we have considered explicitly Markovian spreading processes, our composition and rejection scheme can be directly applied to certain classes of non-Markovian processes with completely monotone survival function, using the Laplace transform, in the spirit of Ref. [192]. It can also directly be used with a variety of spreading processes on time-varying networks, where the structure evolves independently from the dynamics [268, 274], or co-evolves with it [114, 249]. Extension of this method for complex contagion [176] would further be an interesting avenue.

Finally, from a more general perspective, we can argue that the idea behind composition and rejection is a systematic and efficient regrouping of independent processes, especially suited for multiscale propensities, as discussed in Sec. A.6.2. It would be simple to exploit this scheme for many of the stochastic processes studied in network science, such as multistate dynamical processes [85, 106] or network generation models [119, 166].

## A.9 Appendix : Some results for uncorrelated random networks

Over the last two decades, many analytical results have been obtained for spreading processes on uncorrelated random networks with an arbitrary degree distribution  $P(k)$ , i.e. the probability of having

---

3. A larger dispersion for the degree distribution implies a more heterogeneous network.

an edge with endpoint nodes of degree  $k, k'$  is  $P(k, k') \propto kk'P(k)P(k')$ . To generate such uncorrelated networks, one must respect the structural cut-off [47], i.e.  $k_{\max} < \langle k \rangle N^{1/2}$ .

In the infinite size limit, threshold estimates  $\lambda_c$  are obtained for spreading processes. For the SIS model, a good approximation is [268]

$$\lambda_c^{\text{SIS}} = \min \left( \sqrt{\frac{2}{k_{\max}}}, \frac{\langle k \rangle}{\langle k^2 \rangle - \langle k \rangle} \right), \quad (\text{A.24})$$

while for the SIR model, the exact result is [201, 284]

$$\lambda_c^{\text{SIR}} = \frac{\langle k \rangle}{\langle k^2 \rangle - 2 \langle k \rangle}. \quad (\text{A.25})$$

Equations (A.24) and (A.25) were used in the simulations of Fig. A.4.

Using the heterogeneous mean-field theory [37, 209], we can also approximate the degree distribution of infected nodes near the phase transition. On the one hand, for the SIS model in the stationary state, we have

$$P(k|I)^* \approx \frac{\lambda \theta^* k}{\mu + \lambda \theta^* k} P(k), \quad (\text{A.26})$$

with degree distribution  $P(k)$ , and where  $\theta^*$  is the average fraction of infected neighbors for a susceptible node. Near the phase transition,  $\theta^* \rightarrow 0$ , which means that the average degree of infected nodes is

$$\langle k|I \rangle^* \sim \langle k^2 \rangle. \quad (\text{A.27})$$

On the other hand, if we consider a complete sequence of the SIR model, the average degree of recovered nodes is a good proxy, which is also  $\sim \langle k^2 \rangle$  near the phase transition. For power-law degree distribution  $P(k) \propto k^{-\gamma}$ , maximal degree  $k_{\max} < \langle k \rangle N^{1/2}$  and  $2 \leq \gamma < 3$ , the second moment of the degree distribution scales with the number of nodes as

$$\langle k^2 \rangle \sim \int_{k_{\min}}^{k_{\max}} k^{2-\gamma} \sim N^{(3-\gamma)/2}. \quad (\text{A.28})$$

## A.10 Appendix : Overhead factor due to phantom events

An element that we have discarded in our complexity analysis is the average fraction of *phantom processes* [61]  $\phi$ , i.e. processes  $u_\alpha$  that do not change the state  $\mathbf{x}$ . One example is the transmission to an already infected node. It is assumed in our analysis of Sec. A.6 that  $\phi$  is upper-bounded by a constant  $C < 1$ , independent of the number of nodes in the system—this is well supported by the results of the composition and rejection scheme in Fig. A.4. It is worth pointing that the results of Fig. A.4 are unbiased, since we counted the number of real transitions  $\mathbf{x} \mapsto \mathbf{x}'$  with  $\mathbf{x}' \neq \mathbf{x}$  to evaluate the empirical average computation time.

For different values of prevalence (SIS) or final size (SIR), phantom processes can lead to a certain overhead. We show in Fig. A.5 the expected multiplicative factor on the number of operations required for an update of the state, due to a certain fraction  $\phi$  of phantom processes.

Near the phase transition, the overhead factor is negligible, but it can get important for prevalence or final size near 1 where infected nodes are mostly surrounded by infected or recovered nodes. It would be possible to reduce the fraction of phantom processes  $\phi$  in the node-based methods for the SIR model : we could count the number  $\tilde{k}_i \leq k_i$  of susceptible neighbors for a newly infected node, and modify the propensity to

$$\tilde{\omega}_i = \delta(x_i, I)(\lambda \tilde{k}_i + \mu). \quad (\text{A.29})$$

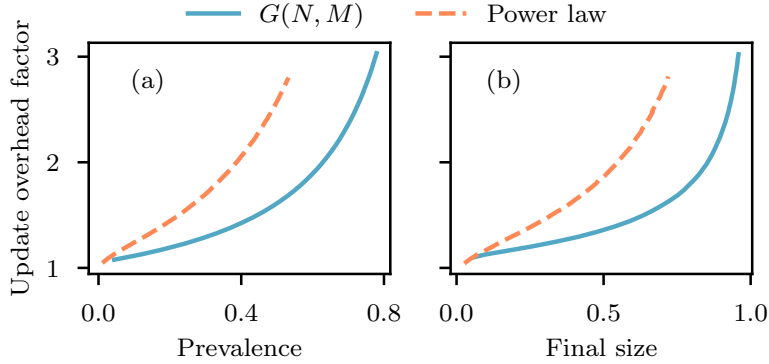


FIGURE A.5 – Expected overhead factor for the update time of a state due to phantom events, using node-based methods. This factor is estimated by  $1/(1 - \phi)$  and  $\phi$  is measured for (a) the SIS model in the stationary state and (b) complete sequence of the SIR model. We sampled over 10 realizations of random graphs ensembles : the  $G(N, M)$  ensemble with  $N = 10^5$  and  $M = 5 \times 10^5$ ; random graphs with an expected degree sequence, a power law expected degree distribution  $P(\kappa) \propto \kappa^{-\gamma}$  with  $\gamma = 2.5$ , a natural cut-off  $k_{\max} < N$  and  $N = 10^5$ .

However, the update involving the infection of a degree  $k_I$  node would now require

$$O[\log(\log N) + k_I], \quad (\text{A.30})$$

which could be worst in certain cases according to our study. Since  $O[\log(\log N)]$ —or  $O(1)$  for upper-bounded maximal degree—is a small number of operations, it is probably safer (and simpler) to keep the schemes introduced in Sec. A.6.2.

## A.11 Supplemental Material

### A.11.1 Computation time comparison without rescaling

To facilitate the comparison of various methods across different programming languages, we have rescaled by a constant factor the computation time of the event-driven method (Fig. A.4 in the main text). Figure A.6 shows a comparison of the methods without rescaling.

### A.11.2 Additional computation time comparison for the SIR model

The decision to compare our method with the package of Ref. [163] is mainly due to the fact they provide a user-friendly interface, both for the SIS and the SIR. However, the comparison across programming languages hinders the absolute (true) computation time. Figure A.7 provides a comparison of the computation time between our method and that of Petter Holme [132], an event-driven algorithm for the SIR model. We can see that both methods are competitive, ours being preferable for dense or large and heterogeneous networks.

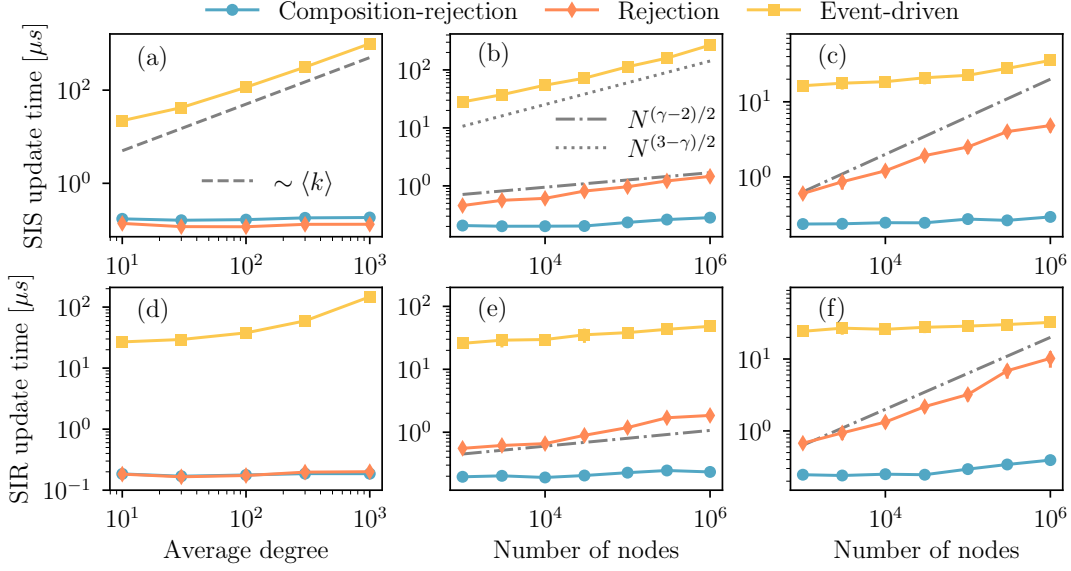


FIGURE A.6 – Average computation time for a single update of the state for spreading processes using different algorithms. Each marker is averaged over 10 simulations on 10 realizations of a random graph ensemble. The rejection and composition and rejection methods are implemented in C++, while the event-driven method is implemented in Python. (Upper row) Average over sequences of  $10^6$  state transitions for the SIS model in the stationary state. The systems had been thermalized using  $10^6$  transitions beforehand. (Lower row) Average over  $10^6/N$  complete sequences of the SIR model, starting with an initial infected node density of  $10^{-2}$ . (a) and (d)  $G(N, M)$  random graphs with fixed number of nodes  $N = 10^4$  and different average degree  $\langle k \rangle$ . The recovery and transmission rates used are  $\mu = 1$  and  $\lambda = 1.05\lambda_c$ . (b) and (e) Random graphs with an expected degree sequence [57, 203]. We used a power law expected degree distribution  $P(\kappa) \propto \kappa^{-\gamma}$  with  $\gamma = 2.25$ ,  $\kappa_{\min} = 3$  and  $\kappa_{\max} < \langle k \rangle N^{1/2}$ . The recovery and transmission rates used are  $\mu = 1$  and  $\lambda = 3\lambda_c$ . (c) and (f) Same as (b) and (e), but with  $\gamma = 3$ .

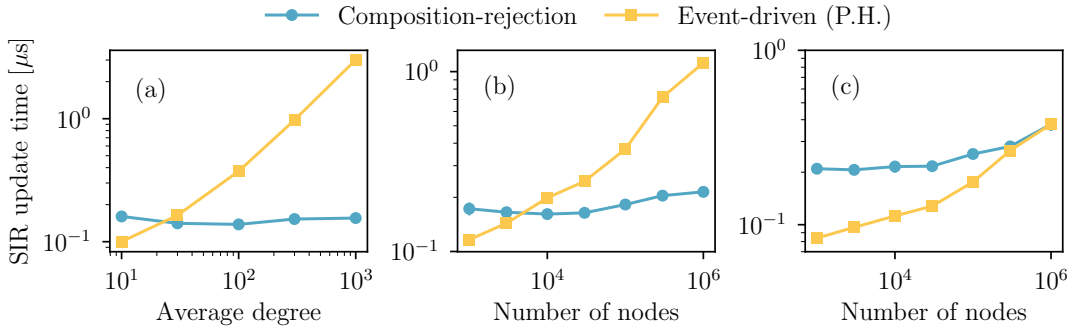


FIGURE A.7 – Average computation time for a single update of the state for the SIR model using different algorithms. Each marker is averaged over 10 simulations on 10 realizations of a random graph ensemble. The composition and rejection method is implemented in C++ and the event-driven method of Ref. [132] is implemented in C. Averages are performed over  $10^6/N$  complete sequences of the SIR model, starting with an initial infected node density of  $10^{-2}$ . (a)  $G(N, M)$  random graphs with fixed number of nodes  $N = 10^4$  and different average degree  $\langle k \rangle$ . The recovery and transmission rates used are  $\mu = 1$  and  $\lambda = 1.05\lambda_c$ . (b) Random graphs with an expected degree sequence [57, 203]. We used a power law expected degree distribution  $P(\kappa) \propto \kappa^{-\gamma}$  with  $\gamma = 2.25$ ,  $\kappa_{\min} = 3$  and  $\kappa_{\max} < \langle k \rangle N^{1/2}$ . The recovery and transmission rates used are  $\mu = 1$  and  $\lambda = 3\lambda_c$ . (c) Same as (b), but with  $\gamma = 3$ .



# Bibliographie

- [1] *DBLP*. <https://dblp.org/xml/release/>. Accessed : 2021-08-23.
- [2] *SocioPatterns*. <http://www.sociopatterns.org/datasets/>. Accessed : 2021-08-23.
- [3] *COVID-19 source investigation*. <https://www.cdc.gov/coronavirus/2019-ncov/php/contact-tracing/contact-tracing-plan/source-investigation.html> (2021). Accessed : 2021-06-25.
- [4] ABBE, E. *Community detection and stochastic block models : recent developments*. *J. Mach. Learn. Res.*, **18** (2017), page 6446.
- [5] ALLARD, A. et HÉBERT-DUFRESNE, L. *Percolation and the effective structure of complex networks*. *Phys. Rev. X*, **9** (2019), page 011023.
- [6] ALLARD, A., HÉBERT-DUFRESNE, L., NOËL, P.-A., MARCEAU, V. et DUBÉ, L. J. *Bond percolation on a class of correlated and clustered random graphs*. *J. Phys. A*, **45** (2012), page 405005.
- [7] ALLARD, A., HÉBERT-DUFRESNE, L., YOUNG, J.-G. et DUBÉ, L. J. *General and exact approach to percolation on random graphs*. *Phys. Rev. E*, **92** (2015), page 062807.
- [8] ALTHOUSE, B. M., WENGER, E. A., MILLER, J. C., SCARPINO, S. V., ALLARD, A., HÉBERT-DUFRESNE, L. et HU, H. *Superspreading events in the transmission dynamics of SARS-CoV-2 : Opportunities for interventions and control*. *PLOS Biol.*, **18** (2020), page e3000897.
- [9] ALVAREZ-RODRIGUEZ, U., BATTISTON, F., FERRAZ DE ARRUDA, G., MORENO, Y., PERC, M. et LATORA, V. *Evolutionary dynamics of higher-order interactions in social networks*. *Nat. Hum. Behav.*, **5** (2021), page 586.
- [10] AMATO, F., MOSCATO, V., PICARIELLO, A. et SPERLÍ, G. *Influence maximization in social media networks using hypergraphs*. Dans *Green, Pervasive, and Cloud Computing*. Springer (2017), page 207.
- [11] ANDERSON, P. W. *Absence of diffusion in certain random lattices*. *Phys. Rev.*, **109** (1958), page 1492.
- [12] ANDERSON, R. M., ANDERSON, B. et MAY, R. M. *Infectious Diseases of Humans : Dynamics and Control*. Oxford University Press (1991).
- [13] ANTTILA, J., MIKONRANTA, L., KETOLA, T., KAITALA, V., LAAKSO, J. et RUOKOLAINEN, L. *A mechanistic underpinning for sigmoid dose-dependent infection*. *Oikos*, **126** (2017), page 910.
- [14] ARENAS, A., COTA, W., GÓMEZ-GARDEÑES, J., GÓMEZ, S., GRANELL, C., MATAMALAS, J. T., SORIANO-PAÑOS, D. et STEINEGGER, B. *Modeling the spatiotemporal epidemic spreading of COVID-19 and the impact of mobility and social distancing interventions*. *Phys. Rev. X*, **10** (2020), page 041055.
- [15] BALASUBRAMANIAN, V., WIEGESHAUS, E. H., TAYLOR, B. T. et SMITH, D. W. *Pathogenesis of tuberculosis : Pathway to apical localization*. *Tuber. Lung Dis.*, **75** (1994), page 168.

- [16] BARABÁSI, A.-L. et ALBERT, R. *Emergence of scaling in random networks*. Science, **286** (1999), page 509.
- [17] BARABÁSI, A.-L. *Love is all you need*. <https://www.barabasilab.com/post/love-is-all-you-need> (2018). Accessed : 2021-07-11.
- [18] BARRAT, A., BARTHÉLEMY, M., PASTOR-SATORRAS, R. et VESPIGNANI, A. *The architecture of complex weighted networks*. Proc. Natl. Acad. Sci. U.S.A., **101** (2004), page 3747.
- [19] BARRAT, A., BARTHÉLEMY, M. et VESPIGNANI, A. *Dynamical Processes on Complex Networks*. Cambridge University Press (2008).
- [20] BARRAT, A., FERRAZ DE ARRUDA, G., IACOPINI, I. et MORENO, Y. *Social contagion on higher-order structures*. arXiv :2103.03709, (2021).
- [21] BARTHEL, T., DE BACCO, C. et FRANZ, S. *Matrix product algorithm for stochastic dynamics on networks applied to nonequilibrium glauber dynamics*. Phys. Rev. E, **97** (2018), page 010104.
- [22] BATTISTON, F., CENCETTI, G., IACOPINI, I., LATORA, V., LUCAS, M., PATANIA, A., YOUNG, J.-G. et PETRI, G. *Networks beyond pairwise interactions : Structure and dynamics*. Phys. Rep., **874** (2020), page 1.
- [23] BEAUCHEMIN, C. A. et HANDEL, A. *A review of mathematical models of influenza A infections within a host or cell culture : lessons learned and challenges ahead*. BMC Public Health, **11** (2011), page S7.
- [24] BENDER, C. M. et ORSZAG, S. A. *Advanced Mathematical Methods for Scientists and Engineers I : Asymptotic Methods and Perturbation Theory*. Springer (1999).
- [25] BENSON, A. R., ABEBE, R., SCHAUB, M. T., JADBABAIE, A. et KLEINBERG, J. *Simplicial closure and higher-order link prediction*. Proc. Natl. Acad. Sci. U.S.A., **115** (2018), page E11221.
- [26] BERDAHL, A. *et al.* *Dynamics of beneficial epidemics*. Sci. Rep., **9** (2019), page 15093.
- [27] BERGE, C. *Hypergraphs : Combinatorics of Finite Sets*. North Holland (1989).
- [28] BERNOULLI, D. *Essai d'une nouvelle analyse de la mortalité causée par la petite vérole*. Mem. Math. Phys. Acad. Roy. Sci., (1766).
- [29] BI, Q. *et al.* *Epidemiology and transmission of COVID-19 in 391 cases and 1286 of their close contacts in shenzhen, china : a retrospective cohort study*. Lancet Infect. Dis., **20** (2020), page 911.
- [30] BIANCONI, G. *Entropy of network ensembles*. Phys. Rev. E, **79** (2009), page 036114.
- [31] BIANCONI, G. et RAHMEDE, C. *Emergent Hyperbolic Network Geometry*. Sci. Rep., **7** (2017), page 41974.
- [32] BICK, C., ASHWIN, P. et RODRIGUES, A. *Chaos in generically coupled phase oscillator networks with nonpairwise interactions*. Chaos, **26** (2016), page 094814.
- [33] BIGGERSTAFF, M. *et al.* *Results from the second year of a collaborative effort to forecast influenza seasons in the United States*. Epidemics, **24** (2018), page 26.
- [34] BLOEM-REDDY, B. et ORBANZ, P. *Random-walk models of network formation and sequential monte carlo methods for graphs*. J. R. Stat. Soc. B, **80** (2018), page 871.
- [35] BOCCALETTI, S. *et al.* *The structure and dynamics of multilayer networks*. Phys. Rep., **544** (2014), page 1.
- [36] BODÓ, Á., KATONA, G. Y. et SIMON, P. L. *SIS epidemic propagation on hypergraphs*. Bull. Math. Biol., **78** (2016), page 713.

- [37] BOGUÑÁ, M. et PASTOR-SATORRAS, R. *Epidemic spreading in correlated complex networks*. Phys. Rev. E, **66** (2002), page 047104.
- [38] BOGUÑÁ, M., PASTOR-SATORRAS, R. et VESPIGNANI, A. *Cut-offs and finite size effects in scale-free networks*. Eur. Phys. J. B, **38** (2004), page 205.
- [39] BROIDO, A. D. et CLAUSET, A. *Scale-free networks are rare*. Nat. Commun., **10** (2019), page 1017.
- [40] BURGIO, G., ARENAS, A., GÓMEZ, S. et MATAMALAS, J. T. *Network clique cover approximation to analyze complex contagions through group interactions*. Commun. Phys., **4** (2021), page 1.
- [41] BUSENBERG, S. N., IANNELLI, M. et THIEME, H. R. *Global behavior of an age-structured epidemic model*. SIAM J. Math. Anal., **22** (1991), page 1065.
- [42] CANTWELL, G. T., ST-ONGE, G. et YOUNG, J.-G. *Inference, model selection, and the combinatorics of growing trees*. Phys. Rev. Lett., **126** (2021), page 038301.
- [43] CARLETTI, T., BATTISTON, F., CENCETTI, G. et FANELLI, D. *Random walks on hypergraphs*. Phys. Rev. E, **101** (2020), page 022308.
- [44] CASADEVALL, A. *Evolution of intracellular pathogens*. Annu. Rev. Microbiol., **62** (2008), page 19.
- [45] CASTELLANO, C. et PASTOR-SATORRAS, R. *Competing activation mechanisms in epidemics on networks*. Sci. Rep., **2** (2012), page 371.
- [46] CASTELLANO, C. et PASTOR-SATORRAS, R. *Relating topological determinants of complex networks to their spectral properties : Structural and dynamical effects*. Phys. Rev. X, **7** (2017), page 41024.
- [47] CATANZARO, M., BOGUÑÁ, M. et PASTOR-SATORRAS, R. *Generation of uncorrelated random scale-free networks*. Phys. Rev. E, **71** (2005), page 027103.
- [48] CATTUTO, C., VAN DEN BROECK, W., BARRAT, A., COLIZZA, V., PINTON, J.-F. et VESPIGNANI, A. *Dynamics of person-to-person interactions from distributed RFID sensor networks*. PLOS ONE, **5** (2010), page e11596.
- [49] CENCETTI, G., BATTISTON, F., LEPRI, B. et KARSAI, M. *Temporal properties of higher-order interactions in social networks*. Sci. Rep., **11** (2021), page 7028.
- [50] CENTOLA, D. *The spread of behavior in an online social network experiment*. Science, **329** (2010), page 1194.
- [51] CENTOLA, D. *How Behavior Spreads : The Science of Complex Contagions*. Princeton University Press (2018).
- [52] CENTOLA, D. et MACY, M. *Complex contagions and the weakness of long ties*. Am. J. Sociol., **113** (2007), page 702.
- [53] CHEN, D., LÜ, L., SHANG, M.-S., ZHANG, Y.-C. et ZHOU, T. *Identifying influential nodes in complex networks*. Physica A, **391** (2012), page 1777.
- [54] CHEN, W., YUAN, Y. et ZHANG, L. *Scalable influence maximization in social networks under the linear threshold model*. Dans *2010 IEEE international conference on data mining*. IEEE (2010), page 88.
- [55] CHODROW, P. S. *Configuration models of random hypergraphs*. J. Complex Netw., **8** (2020), page cnaa018.
- [56] CHOWDHARY, S., KUMAR, A., CENCETTI, G., IACOPINI, I. et BATTISTON, F. *Simplicial contagion in temporal higher-order networks*. J. Phys. Complex., **2** (2021), page 035019.

- [57] CHUNG, F. et LU, L. *Connected components in random graphs with given expected degree sequences*. Ann. Comb., **6** (2002), page 125.
- [58] CIMINI, G., SQUARTINI, T., SARACCO, F., GARLASCHELLI, D., GABRIELLI, A. et CALDARELLI, G. *The statistical physics of real-world networks*. Nat. Rev. Phys., **1** (2019), page 58.
- [59] CISNEROS-VELARDE, P. et BULLO, F. *Multi-group SIS epidemics with simplicial and higher-order interactions*. arXiv :2005.11404, (2021).
- [60] COLIZZA, V., BARRAT, A., BARTHÉLEMY, M. et VESPIGNANI, A. *The role of the airline transportation network in the prediction and predictability of global epidemics*. Proc. Natl. Acad. Sci. U.S.A., **103** (2006), page 2015.
- [61] COTA, W. et FERREIRA, S. C. *Optimized Gillespie algorithms for the simulation of markovian epidemic processes on large and heterogeneous networks*. Comput. Phys. Commun., **219** (2017), page 303.
- [62] COTA, W., FERREIRA, S. C. et ÓDOR, G. *Griffiths effects of the susceptible-infected-susceptible epidemic model on random power-law networks*. Phys. Rev. E, **93** (2016), page 032322.
- [63] COTA, W., MATA, A. S. et FERREIRA, S. C. *Robustness and fragility of the susceptible-infected-susceptible epidemic models on complex networks*. Phys. Rev. E, **98** (2018), page 012310.
- [64] COTA, W., ÓDOR, G. et FERREIRA, S. C. *Griffiths phases in infinite-dimensional, non-hierarchical modular networks*. Sci. Rep., **8** (2018), page 9144.
- [65] COURTNEY, O. T. et BIANCONI, G. *Weighted growing simplicial complexes*. Phys. Rev. E, **95** (2017), page 062301.
- [66] DALEY, D. J. et KENDALL, D. G. *Epidemics and rumours*. Nature, **204** (1964), page 1118.
- [67] DALEY, D. J. et KENDALL, D. G. *Stochastic rumours*. IMA J. Appl. Math., **1** (1965), page 42.
- [68] DE OLIVEIRA, M. M. et DICKMAN, R. *How to simulate the quasistationary state*. Phys. Rev. E, **71** (2005), page 016129.
- [69] DEL GENIO, C. I., GROSS, T. et BASSLER, K. E. *All scale-free networks are sparse*. Phys. Rev. Lett., **107** (2011), page 178701.
- [70] DEVROYE, L. *Non-Uniform Random Variate Generation*. Springer (1986).
- [71] DI LAURO, F., CROIX, J.-C., BERTHOUBE, L. et KISS, I. Z. *PDE limits of stochastic SIS epidemics on networks*. J. Complex Netw., **8** (2020), page cnaa043.
- [72] DIEKMANN, O., METZ, J. A. J. et HEESTERBEEK, J. A. P. *The legacy of Kermack and McKendrick*. Dans *Epidemic Models : Their Structure and Relation to Data*, édité par MOLLISON, D. Cambridge University Press (1995), page 95.
- [73] DODDS, P. et WATTS, D. J. *Threshold models of social influence*. Dans *The Oxford Handbook of Analytical Sociology*, édité par BEARMAN, P. et HEDSTRÖM, P. Oxford University Press (2009), page 475.
- [74] DODDS, P. S. *Slightly generalized contagion : Unifying simple models of biological and social spreading*. Dans *Complex Spreading Phenomena in Social Systems*, édité par LEHMANN, S. et AHN, Y.-Y. Springer (2018), page 67.
- [75] DODDS, P. S. et WATTS, D. J. *Universal behavior in a generalized model of contagion*. Phys. Rev. Lett., **92** (2004), page 218701.
- [76] DOMINGOS, P. et RICHARDSON, M. *Mining the network value of customers*. Dans *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery (2001), page 57.

- [77] DU, S. Q. et YUAN, W. *Mathematical modeling of interaction between innate and adaptive immune responses in COVID-19 and implications for viral pathogenesis*. J. Med. Virol., **92** (2020), page 1615.
- [78] EAMES, K. T. D. et KEELING, M. J. *Modeling dynamic and network heterogeneities in the spread of sexually transmitted diseases*. Proc. Natl. Acad. Sci. USA, **99** (2002), page 13330.
- [79] ENDO, A., ABBOTT, S., KUCHARSKI, A. J. et FUNK, S. *Estimating the overdispersion in COVID-19 transmission using outbreak sizes outside china*. Wellcome Open Res., **5** (2020), page 67.
- [80] EPSTEIN, J. M. *Why Model?* J. Artif. Soc. Soc. Simul., **11** (2008), page 12.
- [81] ERDÖS, P. et RÉNYI, A. *On random graphs I*. Publ. Math. Debr., **6** (1959), page 290.
- [82] ERDÖS, P. L., MIKLÓS, I. et SOUKUP, L. *Towards random uniform sampling of bipartite graphs with given degree sequence*. Electron. J. Comb., **20** (2013), page 16.
- [83] ERKOL, Ş., CASTELLANO, C. et RADICCHI, F. *Systematic comparison between methods for the detection of influential spreaders in complex networks*. Sci. Rep., **9** (2019), page 15095.
- [84] FENG, Z. et THIEME, H. R. *Endemic models with arbitrarily distributed periods of infection  $i$  : Fundamental properties of the model*. SIAM J. Appl. Math., **61** (2000), page 803.
- [85] FENNEL, P. G. et GLEESON, J. P. *Multistate dynamical processes on networks : analysis through degree-based approximation frameworks*. SIAM Rev., **61** (2019), page 92.
- [86] FENNEL, P. G., MELNIK, S. et GLEESON, J. P. *Limitations of discrete-time approaches to continuous-time contagion dynamics*. Phys. Rev. E, **94** (2016), page 052125.
- [87] FERRAZ DE ARRUDA, G., COZZO, E., PEIXOTO, T. P., RODRIGUES, F. A. et MORENO, Y. *Disease localization in multilayer networks*. Phys. Rev. X, **7** (2017), page 011014.
- [88] FERRAZ DE ARRUDA, G., PETRI, G. et MORENO, Y. *Social contagion models on hypergraphs*. Phys. Rev. Research, **2** (2020), page 023032.
- [89] FERRAZ DE ARRUDA, G., RODRIGUES, F. A. et MORENO, Y. *Fundamentals of spreading processes in single and multilayer complex networks*. Phys. Rep., **756** (2018), page 1.
- [90] FERRAZ DE ARRUDA, G., TIZZANI, M. et MORENO, Y. *Phase transitions and stability of dynamical processes on hypergraphs*. Commun. Phys., **4** (2021), page 24.
- [91] FERREIRA, S. C., SANDER, R. S. et PASTOR-SATORRAS, R. *Collective versus hub activation of epidemic phases on networks*. Phys. Rev. E, **93** (2016), page 32314.
- [92] FINLAY, B. B. et FALKOW, S. *Common themes in microbial pathogenicity revisited*. Microbiol. Mol. Biol. Rev., **61** (1997), page 136.
- [93] FORTUNATO, S. *Community detection in graphs*. Phys. Rep., **486** (2010), page 75.
- [94] FORTUNATO, S. et HRIC, D. *Community detection in networks : A user guide*. Phys. Rep., **659** (2016), page 1.
- [95] FOSDICK, B. K., LARREMORE, D. B., NISHIMURA, J. et UGANDER, J. *Configuring random graph models with fixed degree sequences*. SIAM Rev., **60** (2018), page 315.
- [96] GALE, D. *A theorem on flows in networks*. Pac. J. Math., **7** (1957), page 1073.
- [97] GAMA, J. A., ABBY, S. S., VIEIRA-SILVA, S., DIONISIO, F. et ROCHA, E. P. C. *Immune subversion and quorum-sensing shape the variation in infectious dose among bacterial pathogens*. PLOS Pathog., **8** (2012), page e1002503.

- [98] GARDINER, C. *Stochastic Methods*. Springer (1985).
- [99] GELMAN, A., CARLIN, J. B., STERN, H. S. et RUBIN, D. B. *Bayesian data analysis*. Chapman and Hall/CRC (1995).
- [100] GÉNOIS, M. et BARRAT, A. *Can co-location be used as a proxy for face-to-face contacts?* EPJ Data Sci., **7** (2018), page 11.
- [101] GIBSON, M. A. et BRUCK, J. *Efficient exact stochastic simulation of chemical systems with many species and many channels*. J. Phys. Chem. A, **104** (2000), page 1876.
- [102] GILLESPIE, D. T. *A general method for numerically simulating the stochastic time evolution of coupled chemical reactions*. J. Comput. Phys., **22** (1976), page 403.
- [103] GIRVAN, M. et NEWMAN, M. E. J. *Community structure in social and biological networks*. Proc. Natl. Acad. Sci. U. S. A., **99** (2002), page 7821.
- [104] GLEESON, J. P. *Cascades on correlated and modular random networks*. Phys. Rev. E, **77** (2008), page 046117.
- [105] GLEESON, J. P. *High-accuracy approximation of binary-state dynamics on networks*. Phys. Rev. Lett., **107** (2011), page 068701.
- [106] GLEESON, J. P. *Binary-state dynamics on complex networks : Pair approximation and beyond*. Phys. Rev. X, **3** (2013), page 021004.
- [107] GLEESON, J. P. et PORTER, M. A. *Message-passing methods for complex contagions*. Dans *Complex Spreading Phenomena in Social Systems*, édité par LEHMANN, S. et AHN, Y.-Y. Springer (2018), page 81.
- [108] GOFFMAN, W. et NEWILL, V. A. *Generalization of epidemic theory : an application to the transmission of ideas*. Nature, **204** (1964), page 225.
- [109] GOLTSEV, A. V., DOROGOVTSSEV, S. N., OLIVEIRA, J. G. et MENDES, J. F. F. *Localization and spreading of diseases in complex networks*. Phys. Rev. Lett., **109** (2012), page 128702.
- [110] GÓMEZ, S., ARENAS, A., BORGE-HOLTHOEFER, J., MELONI, S. et MORENO, Y. *Discrete-time markov chain approach to contact-based disease spreading in complex networks*. EPL, **89** (2010), page 38009.
- [111] GOUTSIAS, J. et JENKINSON, G. *Markovian dynamics on complex reaction networks*. Phys. Rep., **529** (2013), page 199.
- [112] GRANOVETTER, M. *Threshold Models of Collective Behavior*. Am. J. Sociol., **83** (1978), page 1420.
- [113] GRILLI, J., BARABÁS, G., MICHALSKA-SMITH, M. J. et ALLESINA, S. *Higher-order interactions stabilize dynamics in competitive network models*. Nature, **548** (2017), page 210.
- [114] GROSS, T., D’LIMA, C. J. D. et BLASIUS, B. *Epidemic dynamics on an adaptive network*. Phys. Rev. Lett., **96** (2006), page 208701.
- [115] GROSS, T. et SAYAMA, H., éditeurs. *Adaptive Networks*. Springer (2009).
- [116] GROSSMANN, G. et WOLF, V. *Rejection-based simulation of stochastic spreading processes on complex networks*. Dans *International Workshop on Hybrid Systems Biology*. Springer (2019), page 63.
- [117] HATCHER, A. *Algebraic Topology*. Cambridge University Press (2002).
- [118] HÉBERT-DUFRESNE, L. et ALLARD, A. *Smearred phase transitions in percolation on real complex networks*. Phys. Rev. Research, **1** (2019), page 013009.

- [119] HÉBERT-DUFRESNE, L., ALLARD, A., MARCEAU, V., NOËL, P.-A. et DUBÉ, L. J. *Structural preferential attachment : Network organization beyond the link*. Phys. Rev. Lett., **107** (2011), page 158702.
- [120] HÉBERT-DUFRESNE, L., ALLARD, A., YOUNG, J.-G. et DUBÉ, L. J. *Global efficiency of local immunization on complex networks*. Sci. Rep., **3** (2013), page 2171.
- [121] HÉBERT-DUFRESNE, L., ALLARD, A., YOUNG, J.-G. et DUBÉ, L. J. *Constrained growth of complex scale-independent systems*. Phys. Rev. E, **93** (2016), page 032304.
- [122] HÉBERT-DUFRESNE, L. et ALTHOUSE, B. M. *Complex dynamics of synergistic coinfections on realistically clustered networks*. Proc. Natl. Acad. Sci. U. S. A., **112** (2015), page 10551.
- [123] HÉBERT-DUFRESNE, L., GROCHOW, J. A. et ALLARD, A. *Multi-scale structure and topological anomaly detection via a new network statistic : The onion decomposition*. Sci. Rep., **6** (2016), page 31708.
- [124] HÉBERT-DUFRESNE, L., LAURENCE, E., ALLARD, A., YOUNG, J.-G. et DUBÉ, L. J. *Complex networks as an emerging property of hierarchical preferential attachment*. Phys. Rev. E, **92** (2015), page 062809.
- [125] HÉBERT-DUFRESNE, L., NOËL, P.-A., MARCEAU, V., ALLARD, A. et DUBÉ, L. J. *Propagation dynamics on networks featuring complex topologies*. Phys. Rev. E, **82** (2010), page 036115.
- [126] HÉBERT-DUFRESNE, L., SCARPINO, S. V. et YOUNG, J.-G. *Macroscopic patterns of interacting contagions are indistinguishable from social reinforcement*. Nat. Phys., **16** (2020), page 426.
- [127] HETHCOTE, H. W. *Modeling heterogeneous mixing in infectious disease dynamics*. Dans *Models for Infectious Human Diseases : Their Structure and Relation to Data*, édité par ISHAM, V. et G., M. Cambridge University Press (1996), page 215.
- [128] HETHCOTE, H. W. et VAN ARK, J. W. *Epidemiological models for heterogeneous populations : Proportionate mixing, parameter estimation, and immunization programs*. Math. Biosci., **84** (1987), page 85.
- [129] HETHCOTE, H. W. et VAN DEN DRIESSCHE, P. *Some epidemiological models with nonlinear incidence*. J. Math. Biol., **29** (1991), page 271.
- [130] HODAS, N. O. et LERMAN, K. *The simple rules of social contagion*. Sci. Rep., **4** (2014), page 4343.
- [131] HOLME, P. *Three faces of node importance in network epidemiology : Exact results for small graphs*. Phys. Rev. E, **96** (2017), page 062305.
- [132] HOLME, P. *SIR*. <https://github.com/pholme/sir> (2018).
- [133] HOLME, P., KIM, B. J., YOON, C. N. et HAN, S. K. *Attack vulnerability of complex networks*. Phys. Rev. E, **65** (2002), page 056109.
- [134] HOLME, P. et SARAMÄKI, J. *Temporal networks*. Phys. Rep., **519** (2012), page 97.
- [135] HOLME, P. et SARAMÄKI, J. *Temporal network theory*. Springer (2019).
- [136] HORNEF, M. W., WICK, M. J., RHEN, M. et NORMARK, S. *Bacterial strategies for overcoming host innate and adaptive immune responses*. Nat. immunol., **3** (2002), page 1033.
- [137] HOUSE, T. et KEELING, M. J. *Deterministic epidemic models with explicit household structure*. Math. Biosci., **213** (2008), page 29.
- [138] HOUSE, T., ROSS, J. V. et SIRL, D. *How big is an outbreak likely to be ? methods for epidemic final-size calculation*. Proc. R. Soc. A, **469** (2013), page 20120436.

- [139] HUNTER, R. F., DE LA HAYE, K., MURRAY, J. M., BADHAM, J., VALENTE, T. W., CLARKE, M. et KEE, F. *Social network interventions for health behaviours and outcomes : A systematic review and meta-analysis*. PLOS Med., **16** (2019), page e1002890.
- [140] HÉBERT-DUFRESNE, L., ALTHOUSE, B. M., SCARPINO, S. V. et ALLARD, A. *Beyond  $R_0$  : heterogeneity in secondary infections and probabilistic epidemic forecasting*. J. R. Soc. Interface, **17** (2020), page 20200393.
- [141] IACOPINI, I., PETRI, G., BARONCHELLI, A. et BARRAT, A. *Vanishing size of critical mass for tipping points in social convention*. arXiv :2103.10411, (2021).
- [142] IACOPINI, I., PETRI, G., BARRAT, A. et LATORA, V. *Simplicial models of social contagion*. Nat. Commun., **10** (2019), page 2485.
- [143] JACKSON, S. J., BAILEY, M. et FOUCAULT WELLES, B. *#hashtagactivism : Networks of Race and Gender Justice*. The MIT Press (2020).
- [144] JACKSON, S. J. et FOUCAULT WELLES, B. *Hijacking #myNYPD : Social media dissent and networked counterpublics*. J. Commun., **65** (2015), page 932.
- [145] JAYNES, E. T. *Probability Theory : The Logic of Science*. Cambridge University Press (2003).
- [146] JEONG, H., MASON, S. P., BARABÁSI, A.-L. et OLTVAI, Z. N. *Lethality and centrality in protein networks*. Nature, **411** (2001), page 41.
- [147] JHUN, B., JO, M. et KAHNG, B. *Simplicial SIS model in scale-free uniform hypergraph*. J. Stat. Mech. : Theory Exp., **2019** (2019), page 123207.
- [148] KARIMI, F. et HOLME, P. *Threshold model of cascades in empirical temporal networks*. Physica A, **392** (2013), page 3476.
- [149] KARRER, B. et NEWMAN, M. E. J. *Message passing approach for general epidemic models*. Phys. Rev. E, **82** (2010), page 016101.
- [150] KARRER, B. et NEWMAN, M. E. J. *Stochastic blockmodels and community structure in networks*. Phys. Rev. E, **83** (2011), page 016107.
- [151] KARSAI, M., IÑIGUEZ, G., KASKI, K. et KERTÉSZ, J. *Complex contagion process in spreading of online innovation*. J. R. Soc. Interface, **11** (2014), page 20140694.
- [152] KARSAI, M., JO, H.-H. et KASKI, K. *Bursty Human Dynamics*. Springer (2018).
- [153] KARSAI, M., KASKI, K., BARABÁSI, A.-L. et KERTÉSZ, J. *Universal features of correlated bursty behaviour*. Sci. Rep., **2** (2012), page 397.
- [154] KARSAI, M., KIVELÄ, M., PAN, R. K., KASKI, K., KERTÉSZ, J., BARABÁSI, A.-L. et SARAMÄKI, J. *Small but slow world : How network topology and burstiness slow down spreading*. Phys. Rev. E, **83** (2011), page 025102.
- [155] KEELING, M. J. *The effects of local spatial structure on epidemiological invasions*. Proc. R. Soc. B, **266** (1999), page 859.
- [156] KEELING, M. J. et EAMES, K. T. *Networks and epidemic models*. J. R. Soc. Interface, **2** (2005), page 295.
- [157] KEELING, M. J. et ROHANI, P. *Modeling Infectious Diseases in Humans and Animals*. Princeton University Press (2008).
- [158] KEMPE, D., KLEINBERG, J. et TARDOS, É. *Maximizing the spread of influence through a social network*. Dans *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery (2003), page 137.



- [159] KENAH, E. et ROBINS, J. M. *Second look at the spread of epidemics on networks*. Phys. Rev. E, **76** (2007), page 036113.
- [160] KERMACK, W. O. et MCKENDRICK, A. G. *Contributions to the mathematical theory of epidemics*. Bull. Math. Biol., **53** (1991), page 33.
- [161] KERMACK, W. O. et MCKENDRICK, A. G. *Contributions to the mathematical theory of epidemics. II - the problem of endemicity*. Bull. Math. Biol., **53** (1991), page 57.
- [162] KERMACK, W. O. et MCKENDRICK, A. G. *Contributions to the mathematical theory of epidemics. III - further studies of the problem of endemicity*. Bull. Math. Biol., **53** (1991), page 89.
- [163] KISS, I. Z., MILLER, J. C. et SIMON, P. L. *Mathematics of Epidemics on Networks : From Exact to Approximate Models*. Springer (2017).
- [164] KITSACK, M., GALLOS, L. K., HAVLIN, S., LILJEROS, F., MUCHNIK, L., STANLEY, H. E. et MAKSE, H. A. *Identification of influential spreaders in complex networks*. Nat. Phys., **6** (2010), pages 888.
- [165] KIVELÄ, M., ARENAS, A., BARTHELEMY, M., GLEESON, J. P., MORENO, Y. et PORTER, M. A. *Multilayer networks*. J. Complex Netw., **2** (2014), page 203.
- [166] KRAPIVSKY, P. L., REDNER, S. et LEYVRAZ, F. *Connectivity of growing random networks*. Phys. Rev. Lett., **85** (2000), page 4629.
- [167] KURTZ, T. G. *Solutions of ordinary differential equations as limits of pure jump markov processes*. J. Appl. Prob., **7** (1970), page 49.
- [168] KURTZ, T. G. *Limit theorems for sequences of jump markov processes approximating ordinary differential processes*. J. Appl. Prob., **8** (1971), page 344.
- [169] LAMBIOTTE, R., ROSVALL, M. et SCHOLTES, I. *From networks to optimal higher-order models of complex systems*. Nat. Phys., **15** (2019), page 313.
- [170] LANDRY, N. W. et RESTREPO, J. G. *The effect of heterogeneity on hypergraph contagion models*. Chaos, **30** (2020), page 103117.
- [171] LAROCQUE, R. C. et CALDERWOOD, S. B. *Syndromes of enteric infection*. Dans Mandell, Douglas, and Bennett's Principles and Practice of Infectious Diseases. Elsevier (2015), page 1238.
- [172] LATAPY, M., MAGNIEN, C. et DEL VECCHIO, N. *Basic notions for the analysis of large two-mode networks*. Soc. Netw., **30** (2008), page 31.
- [173] LAU, M. S. Y., GRENFELL, B., THOMAS, M., BRYAN, M., NELSON, K. et LOPMAN, B. *Characterizing superspreading events and age-specific infectiousness of SARS-CoV-2 transmission in Georgia, USA*. Proc. Natl. Acad. Sci. U.S.A., **117** (2020), page 22430.
- [174] LAWYER, G. *Understanding the influence of all nodes in a network*. Sci. Rep., **5** (2015), page 8665.
- [175] LECLERC, Q. J., FULLER, N. M., KNIGHT, L. E., FUNK, S., KNIGHT, G. M. et CMMID COVID-19 WORKING GROUP. *What settings have been linked to SARS-CoV-2 transmission clusters ?* Wellcome Open Res., **5** (2020).
- [176] LEHMANN, S. et AHN, Y.-Y., éditeurs. *Complex Spreading Phenomena in Social Systems*. Springer (2018).
- [177] LI, Z., DENG, Z., HAN, Z., ALFARO-BITTNER, K., BARZEL, B. et BOCCALETTI, S. *Contagion in simplicial complexes*. Chaos Solit. Fractals, **152** (2021), page 111307.

- [178] LINDQUIST, J., MA, J., VAN DEN DRIESSCHE, P. et WILLEBOORDSE, F. H. *Effective degree network disease models*. J. Math. Biol., **62** (2011), page 143.
- [179] LIPSITCH, M. et O'HAGAN, J. J. *Patterns of antigenic diversity and the mechanisms that maintain them*. J. R. Soc. Interface, **4** (2007), page 787.
- [180] LIU, Q. et VAN MIEGHEM, P. *Network localization is unalterable by infections in bursts*. IEEE Trans. Netw. Sci. Eng., **6** (2019), page 983.
- [181] LIU, Q.-H., AJELLI, M., ALETA, A., MERLER, S., MORENO, Y. et VESPIGNANI, A. *Measurability of the epidemic reproduction number in data-driven contact networks*. Proc. Natl. Acad. Sci. U.S.A., **115** (2018), page 12680.
- [182] LIU, W.-M., HETHCOTE, H. W. et LEVIN, S. A. *Dynamical behavior of epidemiological models with nonlinear incidence rates*. J. Math. Biol., **25** (1987), page 359.
- [183] LIU, W.-M., LEVIN, S. A. et IWASA, Y. *Influence of nonlinear incidence rates upon the behavior of SIRS epidemiological models*. J. Math. Biol., **23** (1986), page 187.
- [184] LIU, Y., EGGO, R. M. et KUCHARSKI, A. J. *Secondary attack rate and superspreading events for SARS-CoV-2*. Lancet, **395** (2020), page e47.
- [185] LLOYD-SMITH, J. O., SCHREIBER, S. J., KOPP, P. E. et GETZ, W. M. *Superspreading and the effect of individual variation on disease emergence*. Nature, **438** (2005), page 355.
- [186] MACEDO-FILHO, A., ALVES, G. A., COSTA FILHO, R. N. et ALVES, T. F. A. *Reactivating dynamics for the susceptible-infected-susceptible model : a simple method to simulate the absorbing phase*. J. Stat. Mech., **2018** (2018), page 043208.
- [187] MACKENZIE, D. *How large a gathering is too large during the coronavirus pandemic ?* <https://www.sciencenews.org/article/coronavirus-covid19-social-gathering-size-math-pandemic> (2020). Accessed : 2021-08-31.
- [188] MAIER, B. F. *epipack : An infectious disease modeling package for python*. J. Open Source Softw., **6** (2021), page 3097.
- [189] MARCEAU, V., NOËL, P.-A., HÉBERT-DUFRESNE, L., ALLARD, A. et DUBÉ, L. J. *Adaptive networks : Coevolution of disease and topology*. Phys. Rev. E, **82** (2010), page 036116.
- [190] MARTIN, T., ZHANG, X. et NEWMAN, M. E. J. *Localization and centrality in networks*. Phys. Rev. E, **90** (2014), page 052808.
- [191] MARTINEZ, N. D. *Artifacts or attributes ? Effects of resolution on the Little Rock Lake food web*. Ecol. Monogr., **61** (1991), page 367.
- [192] MASUDA, N. et ROCHA, L. E. *A Gillespie algorithm for non-Markovian stochastic processes*. SIAM Rev., **60** (2018), page 95.
- [193] MATA, A. S., FERREIRA, R. S. et FERREIRA, S. C. *Heterogeneous pair-approximation for the contact process on complex networks*. New J. Phys., **16** (2014), page 053006.
- [194] MATA, A. S. et FERREIRA, S. C. *Pair quenched mean-field theory for the susceptible-infected-susceptible model on complex networks*. EPL, **103** (2013), page 48003.
- [195] MATA, A. S. et FERREIRA, S. C. *Multiple transitions of the susceptible-infected-susceptible epidemic model on complex networks*. Phys. Rev. E, **91** (2015), page 012816.
- [196] MATAMALAS, J. T., GÓMEZ, S. et ARENAS, A. *Abrupt phase transition of epidemic spreading in simplicial complexes*. Phys. Rev. Research, **2** (2020), page 012049.

- [197] MCFARLAND, L. V. et STAMM, W. E. *Review of clostridium difficile—associated diseases*. Am. J. Infect. Control, **14** (1986), page 99.
- [198] MEZARD, M. et MONTANARI, A. *Information, Physics, and Computation*. Oxford University Press (2009).
- [199] MILLÁN, A. P., TORRES, J. J. et BIANCONI, G. *Explosive higher-order kuramoto dynamics on simplicial complexes*. Phys. Rev. Lett., **124** (2020), page 218301.
- [200] MILLER, D. *et al.*. *Full genome viral sequences inform patterns of SARS-CoV-2 spread into and within israel*. Nat. Commun., **11** (2020), page 5518.
- [201] MILLER, J. C. *A note on a paper by Erik Volz : SIR dynamics in random networks*. J. Math. Biol., **62** (2011), page 349.
- [202] MILLER, J. C. *Mathematics-of-epidemics-on-networks*. <https://github.com/springer-math/Mathematics-of-Epidemics-on-Networks> (2018).
- [203] MILLER, J. C. et HAGBERG, A. *Efficient generation of networks with given expected degrees*. Dans *Algorithms and Models for the Web Graph*. Springer (2011), page 115.
- [204] MISLOVE, A., KOPPULA, H. S., GUMMADI, K. P., DRUSCHEL, P. et BHATTACHARJEE, B. *Growth of the flickr social network*. Dans *Proceedings of the First Workshop on Online Social Networks*. Association for Computing Machinery (2008), page 25.
- [205] MISLOVE, A., MARCON, M., GUMMADI, K. P., DRUSCHEL, P. et BHATTACHARJEE, B. *Measurement and analysis of online social networks*. Dans *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*. Association for Computing Machinery (2007), page 29.
- [206] MISTRY, D. *et al.*. *Inferring high-resolution human mixing patterns for disease modeling*. Nat. Commun., **12** (2021), page 323.
- [207] MØNSTED, B., SAPIEŻYŃSKI, P., FERRARA, E. et LEHMANN, S. *Evidence of complex contagion of information in social media : An experiment using Twitter bots*. PLOS ONE, **12** (2017), page e0184148.
- [208] MORENO, Y., NEKOVEE, M. et PACHECO, A. F. *Dynamics of rumor spreading in complex networks*. Phys. Rev. E, **69** (2004), page 066130.
- [209] MORENO, Y., PASTOR-SATORRAS, R. et VESPIGNANI, A. *Epidemic outbreaks in complex heterogeneous networks*. Eur. Phys. J. B, **26** (2002), page 521.
- [210] MORGAN, A. C., ECONOMOU, D. J., WAY, S. F. et CLAUSET, A. *Prestige drives epistemic inequality in the diffusion of scientific ideas*. EPJ Data Sci., **7** (2018), page 40.
- [211] MORONE, F. et MAKSE, H. A. *Influence maximization in complex networks through optimal percolation*. Nature, **524** (2015), page 65.
- [212] MOUW, T. et VERDERY, A. M. *Network sampling with memory : A proposal for more efficient sampling from social networks*. Sociol. Methodol., **42** (2012), page 206.
- [213] MUÑOZ, M. A., JUHÁSZ, R., CASTELLANO, C. et ÓDOR, G. *Griffiths phases on complex networks*. Phys. Rev. Lett., **105** (2010), page 128701.
- [214] NADARAJAH, S. *Exact distribution of the product of m gamma and n Pareto random variables*. J. Comput. Appl. Math., **235** (2011), page 4496.
- [215] NEMATZADEH, A., FERRARA, E., FLAMMINI, A. et AHN, Y.-Y. *Optimal network modularity for information diffusion*. Phys. Rev. Lett., **113** (2014), page 088701.
- [216] NEUHÄUSER, L., MELLOR, A. et LAMBIOTTE, R. *Multibody interactions and nonlinear consensus dynamics on networked systems*. Phys. Rev. E, **101** (2020), page 032310.

- [217] NEWMAN, M. E. J. *Spread of epidemic disease on networks*. Phys. Rev. E, **66** (2002), page 016128.
- [218] NEWMAN, M. E. J. *Properties of highly clustered networks*. Phys. Rev. E, **68** (2003), page 026121.
- [219] NEWMAN, M. E. J. *Network structure from rich but noisy data*. Nat. Phys., **14** (2018), page 542.
- [220] NEWMAN, M. E. J. *Networks*. Oxford University Press (2018).
- [221] NEWMAN, M. E. J., STROGATZ, S. H. et WATTS, D. J. *Random graphs with arbitrary degree distributions and their applications*. Phys. Rev. E, **64** (2001), page 026118.
- [222] NIELSEN, B. F., SIMONSEN, L. et SNEPPEN, K. *COVID-19 superspreading suggests mitigation by social network modulation*. Phys. Rev. Lett., **126** (2021), page 118301.
- [223] ÓDOR, G. *Localization transition, Lifschitz tails, and rare-region effects in network models*. Phys. Rev. E, **90** (2014), page 032110.
- [224] ÓDOR, G., DICKMAN, R. et ÓDOR, G. *Griffiths phases and localization in hierarchical modular networks*. Sci. Rep., **5** (2015), page 14451.
- [225] OLVER, F. *Asymptotics and special functions*. CRC Press (1997).
- [226] O’SULLIVAN, D. J. P., O’KEEFFE, G. J., FENNELL, P. G. et GLEESON, J. P. *Mathematical modeling of complex contagion on clustered networks*. Front. Phys., **3** (2015), page 71.
- [227] OVERGOOR, J., BENSON, A. et UGANDER, J. *Choosing to grow a graph : Modeling network formation as discrete choice*. Dans *The World Wide Web Conference*. Association for Computing Machinery (2019), page 1409.
- [228] PARSHANI, R., CARMİ, S. et HAVLIN, S. *Epidemic threshold for the susceptible-infectious-susceptible model on random networks*. Phys. Rev. Lett., **104** (2010), page 258701.
- [229] PASTOR-SATORRAS, R. et CASTELLANO, C. *Distinct types of eigenvector localization in networks*. Sci. Rep., **6** (2016), page 18847.
- [230] PASTOR-SATORRAS, R. et CASTELLANO, C. *Eigenvector localization in real networks and its implications for epidemic spreading*. J. Stat. Phys., **173** (2018), page 1110.
- [231] PASTOR-SATORRAS, R., CASTELLANO, C., VAN MIEGHEM, P. et VESPIGNANI, A. *Epidemic processes in complex networks*. Rev. Mod. Phys., **87** (2015), page 925.
- [232] PASTOR-SATORRAS, R. et VESPIGNANI, A. *Epidemic dynamics and endemic states in complex networks*. Phys. Rev. E, **63** (2001), page 066117.
- [233] PASTOR-SATORRAS, R. et VESPIGNANI, A. *Epidemic spreading in scale-free networks*. Phys. Rev. Lett., **86** (2001), page 3200.
- [234] PASTOR-SATORRAS, R. et VESPIGNANI, A. *Immunization of complex networks*. Phys. Rev. E, **65** (2002), page 036104.
- [235] PATANIA, A., PETRI, G. et VACCARINO, F. *The shape of collaborations*. EPJ Data Sci., **6** (2017), page 18.
- [236] PEIXOTO, T. P. *Reconstructing networks with unknown and heterogeneous errors*. Phys. Rev. X, **8** (2018), page 041011.
- [237] PERALTA, A. F. et TORAL, R. *Binary-state dynamics on complex networks : Stochastic pair approximation and beyond*. Phys. Rev. Research, **2** (2020), page 043370.

- [238] PETRI, G. et BARRAT, A. *Simplicial activity driven model*. Phys. Rev. Lett., **121** (2018), page 228301.
- [239] POUX-MÉDARD, G., PASTOR-SATORRAS, R. et CASTELLANO, C. *Influential spreaders for recurrent epidemics on networks*. Phys. Rev. Research, **2** (2020), page 023332.
- [240] POWER, E. A. *Discerning devotion : Testing the signaling theory of religion*. Evol. Hum. Behav., **38** (2017), page 82.
- [241] POWER, E. A. *Social support networks and religiosity in rural south india*. Nat. Hum. Behav., **1** (2017), page 0057.
- [242] RIVERS, C. *et al.* *Using “outbreak science” to strengthen the use of models during epidemics*. Nat. Commun., **10** (2019), page 3102.
- [243] ROBINS, G. et ALEXANDER, M. *Small worlds among interlocking directors : Network structure and distance in bipartite graphs*. Comput. Math. Organ. Theory, **10** (2004), page 69.
- [244] ROGERS, E. M. *Diffusion of Innovations*. Simon and Schuster (2010).
- [245] ROSENBLATT, S. F., SMITH, J. A., GAUTHIER, G. R. et HÉBERT-DUFRESNE, L. *Immunization strategies in networks with missing data*. PLOS Comput. Biol., **16** (2020), page 1.
- [246] RYSER, H. J. *Combinatorial properties of matrices of zeros and ones*. Can. J. Math., **9** (1957), page 371.
- [247] SALNIKOV, V., CASSESE, D. et LAMBIOTTE, R. *Simplicial complexes and complex systems*. Eur. J. Phys., **40** (2018), page 014001.
- [248] SANDER, R. S., COSTA, G. S. et FERREIRA, S. C. *Sampling methods for the quasistationary regime of epidemic processes on regular and complex networks*. Phys. Rev. E, **94** (2016), page 042308.
- [249] SCARPINO, S. V., ALLARD, A. et HÉBERT-DUFRESNE, L. *The effect of a prudent adaptive behaviour on disease transmission*. Nat. Phys., **12** (2016), page 1042.
- [250] SCHAUB, M. T., BENSON, A. R., HORN, P., LIPPNER, G. et JADBABAIE, A. *Random walks on simplicial complexes and the normalized hodge 1-Laplacian*. SIAM Rev., **62** (2020), page 353.
- [251] SCHELLING, T. C. *Dynamic models of segregation*. J. Math. Sociol., **1** (1971), page 143.
- [252] SHARKEY, K. J. *Localization of eigenvector centrality in networks with a cut vertex*. Phys. Rev. E, **99** (2019), page 012315.
- [253] SHARKEY, K. J., KISS, I. Z., WILKINSON, R. R. et SIMON, P. L. *Exact equations for SIR epidemics on tree graphs*. Bull. Math. Biol., **77** (2015), page 614.
- [254] SHRESTHA, M., SCARPINO, S. V. et MOORE, C. *Message-passing approach for recurrent-state epidemic models on networks*. Phys. Rev. E, **92** (2015), page 022821.
- [255] SILK, H., DEMIREL, G., HOMER, M. et GROSS, T. *Exploring the adaptive voter model dynamics with a mathematical triple jump*. New J. Phys., **16** (2014), page 093051.
- [256] SIMON, H. A. *On a class of skew distribution functions*. Biometrika, **42** (1955), page 425.
- [257] SKARDAL, P. S. et ARENAS, A. *Abrupt desynchronization and extensive multistability in globally coupled oscillator simplexes*. Phys. Rev. Lett., **122** (2019), page 248301.
- [258] SLEPOY, A., THOMPSON, A. P. et PLIMPTON, S. J. *A constant-time kinetic monte carlo algorithm for simulation of large biochemical reaction networks*. J. Chem. Phys., **128** (2008), page 05B618.

- [259] SOOD, V. et REDNER, S. *Voter model on heterogeneous graphs*. Phys. Rev. Lett., **94** (2005), page 178701.
- [260] ST-ONGE, G. *SamplableSet*. <https://github.com/gstonge/SamplableSet> (2018).
- [261] ST-ONGE, G. *spreading\_cr*. [https://github.com/gstonge/spreading\\_CR](https://github.com/gstonge/spreading_CR) (2018).
- [262] ST-ONGE, G., IACOPINI, I., LATORA, V., BARRAT, A., PETRI, G., ALLARD, A. et HÉBERT-DUFRESNE, L. *Influential groups for seeding and sustaining nonlinear contagion in heterogeneous hypergraphs*. Commun. Phys., **5** (2022), page 25.
- [263] ST-ONGE, G., SUN, H., ALLARD, A., HÉBERT-DUFRESNE, L. et BIANCONI, G. *heterogeneous-exposure-hons*. <https://github.com/gstonge/heterogeneous-exposure-hons> (2021).
- [264] ST-ONGE, G., SUN, H., ALLARD, A., HÉBERT-DUFRESNE, L. et BIANCONI, G. *Superspreading events data*. <https://docs.google.com/spreadsheets/d/1TTX6b00CIDA9StfHcP8Kn809551cbBjn1rCVzg3TCzM/edit?usp=sharing> (2021). Updated : 2021-02-01.
- [265] ST-ONGE, G., SUN, H., ALLARD, A., HÉBERT-DUFRESNE, L. et BIANCONI, G. *Universal nonlinear infection kernel from heterogeneous exposure on higher-order networks*. Phys. Rev. Lett., **127** (2021), page 158301.
- [266] ST-ONGE, G., THIBEAULT, V., ALLARD, A., DUBÉ, L. J. et HÉBERT-DUFRESNE, L. *Master equation analysis of mesoscopic localization in contagion dynamics on higher-order networks*. Phys. Rev. E, **103** (2021), page 032301.
- [267] ST-ONGE, G., THIBEAULT, V., ALLARD, A., DUBÉ, L. J. et HÉBERT-DUFRESNE, L. *Social confinement and mesoscopic localization of epidemics on networks*. Phys. Rev. Lett., **126** (2021), page 098301.
- [268] ST-ONGE, G., YOUNG, J.-G., LAURENCE, E., MURPHY, C. et DUBÉ, L. J. *Phase transition of the susceptible-infected-susceptible dynamics on time-varying configuration model networks*. Phys. Rev. E, **97** (2018), page 022305.
- [269] STARNINI, M., MACHENS, A., CATTUTO, C., BARRAT, A. et PASTOR-SATORRAS, R. *Immuni-zation strategies for epidemic processes in time-varying contact networks*. J. Theor. Biol., **337** (2013), page 89.
- [270] STEHLÉ, J., BARRAT, A. et BIANCONI, G. *Dynamical and bursty interactions in social networks*. Phys. Rev. E, **81** (2010), page 035101.
- [271] STEHLÉ, J. *et al.* *High-resolution measurements of face-to-face contact patterns in a primary school*. PLOS ONE, **6** (2011), page 1.
- [272] SWINKELS, K. *SARS-CoV-2 superspreading events from around the world*. <http://www.superspreadingdatabase.com/> (2020). Accessed : 2020-07-02.
- [273] TANGE, O. *GNU parallel - the command-line power tool*. ;login : The USENIX Magazine, **36** (2011), page 42.
- [274] TAYLOR, M., TAYLOR, T. J. et KISS, I. Z. *Epidemic threshold and control in a dynamic network*. Phys. Rev. E, **85** (2012), page 016103.
- [275] THIBEAULT, V., ST-ONGE, G., DUBÉ, L. J. et DESROSIERS, P. *Threefold way to the dimension reduction of dynamics on networks : An application to synchronization*. Phys. Rev. Research, **2** (2020), page 043215.
- [276] TORRES, J. J. et BIANCONI, G. *Simplicial complexes : Higher-order spectral dimension and dynamics*. J. Phys. Complex., **1** (2020), page 015002.

- [277] TORRES, L., BLEVINS, A. S., BASSETT, D. et ELIASSI-RAD, T. *The why, how, and when of representations for complex systems*. SIAM Rev., **63** (2021), page 435.
- [278] TUDISCO, F. et HIGHAM, D. J. *Node and edge eigenvector centrality for hypergraphs*. arXiv :2101.06215, (2021).
- [279] VAN MIEGHEM, P., OMIĆ, J. et KOOIJ, R. *Virus spread in networks*. IEEE/ACM Trans. Netw., **17** (2009), page 1.
- [280] VAZQUEZ, A., RÁCZ, B., LUKÁCS, A. et BARABÁSI, A.-L. *Impact of non-poissonian activity patterns on spreading processes*. Phys. Rev. Lett., **98** (2007), page 158702.
- [281] VESTERGAARD, C. L. et GÉNOIS, M. *Temporal Gillespie algorithm : Fast simulation of contagion processes on time-varying networks*. PLOS Comput. Biol., **11** (2015), page e1004579.
- [282] VOITALOV, I., VAN DER HOORN, P., VAN DER HOFSTAD, R. et KRIOUKOV, D. *Scale-free networks well done*. Phys. Rev. Research, **1** (2019), page 033034.
- [283] VOJTA, T. *Rare region effects at classical, quantum and nonequilibrium phase transitions*. J. Phys. A, **39** (2006), page R143.
- [284] VOLZ, E. *SIR dynamics in random networks with heterogeneous connectivity*. J. Math. Biol., **56** (2008), page 293.
- [285] VON NEUMANN, J. *Various techniques used in connection with random digits*. Dans *Monte Carlo Method*, édité par HOUSEHOLDER, A. S., FORSYTHE, G. E. et GERMOND, H. H. US Government Printing Office (1951), page 36.
- [286] WANG, W., TANG, M., STANLEY, H. E. et BRAUNSTEIN, L. A. *Unification of theoretical approaches for epidemic spreading on complex networks*. Rep. Prog. Phys., **80** (2017), page 036603.
- [287] WASSERMAN, S. et FAUST, K. *Social Network Analysis : Methods and Applications*. Cambridge University Press (1994).
- [288] WATTS, D. J. *A simple model of global cascades on random networks*. Proc. Natl. Acad. Sci. U.S.A., **99** (2002), page 5766.
- [289] WATTS, D. J. et DODDS, P. S. *Influentials, networks, and public opinion formation*. J. Consum. Res., **34** (2007), page 441.
- [290] WATTS, D. J., MUHAMAD, R., MEDINA, D. C. et DODDS, P. S. *Multiscale, resurgent epidemics in a hierarchical metapopulation model*. Proc. Natl. Acad. Sci. U.S.A., **102** (2005), page 11157.
- [291] WEAVER, W. *Science and complexity*. Am. Sci., **36** (1948), page 536.
- [292] WEBER, T. P. et STILIANAKIS, N. I. *Inactivation of influenza A viruses in the environment and modes of transmission : A critical review*. J. Infect., **57** (2008), page 361.
- [293] WENG, L., MENCZER, F. et AHN, Y.-Y. *Predicting successful memes using network and community structure*. Dans *Eighth international AAAI conference on weblogs and social media*. AAAI (2014), page 535.
- [294] WILLIAMSON, S. A. et TEC, M. *Random clique covers for graphs with local density and global sparsity*. Dans *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*. PMLR (2020), page 228.
- [295] WONG, F. et COLLINS, J. J. *Evidence that coronavirus superspreading is fat-tailed*. Proc. Natl. Acad. Sci. U.S.A., **117** (2020), page 29416.

- [296] YANG, D., QU, B., YANG, J. et CUDRE-MAUROUX, P. *Revisiting user mobility and social relationships in lbsns : A hypergraph embedding approach*. Dans *The World Wide Web Conference*. Association for Computing Machinery (2019), page 2147.
- [297] YATES, C. A. et KLINGBEIL, G. *Recycling random numbers in the stochastic simulation algorithm*. *J. Chem. Phys.*, **138** (2013), page 094103.
- [298] YOUNG, J.-G., CANTWELL, G. T. et NEWMAN, M. E. J. *Bayesian inference of network structure from unreliable data*. *J. Complex Netw.*, **8** (2021), page cnaa046.
- [299] YOUNG, J.-G., HÉBERT-DUFRESNE, L., ALLARD, A. et DUBÉ, L. J. *Growing networks of overlapping communities with internal structure*. *Phys. Rev. E*, **94** (2016), page 022317.
- [300] YOUNG, J.-G., PETRI, G. et PEIXOTO, T. P. *Hypergraph reconstruction from network data*. *Commun. Phys.*, **4** (2021), page 135.
- [301] YOUNG, J.-G., ST-ONGE, G., DESROSIERS, P. et DUBÉ, L. J. *Universality of the stochastic block model*. *Phys. Rev. E*, **98** (2018), page 032309.
- [302] YOUNG, J.-G., ST-ONGE, G., LAURENCE, E., MURPHY, C., HÉBERT-DUFRESNE, L. et DESROSIERS, P. *Phase transition in the recoverability of network history*. *Phys. Rev. X*, **9** (2019), page 041056.
- [303] ZACHARY, W. W. *An information flow model for conflict and fission in small groups*. *J. Anthropol. Res.*, **33** (1977), page 452.
- [304] ZHAO, K., STEHLÉ, J., BIANCONI, G. et BARRAT, A. *Social network dynamics of face-to-face interactions*. *Phys. Rev. E*, **83** (2011), page 056109.
- [305] ZHU, J., ZHU, J., GHOSH, S., WU, W. et YUAN, J. *Social influence maximization in hypergraph in social networks*. *IEEE Transactions on Network Science and Engineering*, **6** (2019), page 801.
- [306] ZWITTER, A. *Big data ethics*. *Big Data Soc.*, **1** (2014), page 2053951714559253.