UNIVERSITÉ LAVAL

ON THE GROWTH AND STRUCTURE OF SOCIAL SYSTEMS FOLLOWING PREFERENTIAL ATTACHMENT

Thèse

Laurent Hébert-Dufresne

Doctorat en physique
Philosophiæ doctor (Ph.D.)

Québec, Canada

Résumé

L'inégalité est une caractéristique notoire des systèmes sociaux. Dans cette thèse, nous nous attarderons à la distribution et à la structure de la répartition de leurs ressources et activités. Dans ce contexte, leurs extrêmes iniquités tendent à suivre une propriété universelle, l'indépendance d'échelle, qui se manifeste par l'absence d'échelle caractéristique. En physique, les organisations indépendantes d'échelle sont bien connues en théorie des transitions de phase dans laquelle on les observe à des points critiques précis. Ceci suggère que des mécanismes bien définis sont potentiellement responsables de l'indépendance d'échelle des systèmes sociaux. Cette analogie est donc au coeur de cette thèse, dont le but est d'aborder ce problème de nature multidisciplinaire avec les outils de la physique statistique.

En premier lieu, nous montrons qu'un système dont la distribution de ressource croît vers l'indépendance d'échelle se trouve assujetti à deux contraintes temporelles particulières. La première est l'attachement préférentiel, impliquant que les riches s'enrichissent. La seconde est une forme générale de comportement d'échelle à délai entre la croissance de la population et celle de la ressource. Ces contraintes dictent un comportement si précis qu'une description instantanée d'une distribution est suffisante pour reconstruire son évolution temporelle et prédire ses états futurs. Nous validons notre approche au moyen de diverses sphères d'activités humaines dont les activités des utilisateurs d'une page web, des relations sexuelles dans une agence d'escorte, ainsi que la productivité d'artistes et de scientifiques.

En second lieu, nous élargissons notre théorie pour considérer la structure résultante de ces activités. Nous appliquons ainsi nos travaux à la théorie des réseaux complexes pour décrire la structure des connexions propre aux systèmes sociaux. Nous proposons qu'une importante classe de systèmes complexes peut être modélisée par une construction hiérarchique de niveaux d'organisation suivant notre théorie d'attachement préférentiel. Nous montrons comment les réseaux complexes peuvent être interprétés comme une projection de ce modèle de laquelle émerge naturellement non seulement leur indépendance d'échelle, mais aussi leur modularité, leur structure hiérarchique, leurs caractéristiques fractales et leur navigabilité. Nos résultats suggèrent que les réseaux sociaux peuvent être relativement simples, et que leur complexité apparente est largement une réflexion de la structure hiérarchique complexe de notre monde.

Abstract

Social systems are notoriously unfair. In this thesis, we focus on the distribution and structure of shared resources and activities. Through this lens, their extreme inequalities tend to roughly follow a universal pattern known as scale independence which manifests itself through the absence of a characteristic scale. In physical systems, scale-independent organizations are known to occur at critical points in phase transition theory. The position of this critical behaviour being very specific, it is reasonable to expect that the distribution of a social resource might also imply specific mechanisms. This analogy is the basis of this work, whose goal is to apply tools of statistical physics to varied social activities.

As a first step, we show that a system whose resource distribution is growing towards scale independence is subject to two constraints. The first is the well-known preferential attachment principle, a mathematical principle roughly stating that the rich get richer. The second is a new general form of delayed temporal scaling between the population size and the amount of available resource. These constraints pave a precise evolution path, such that even an instantaneous snapshot of a distribution is enough to reconstruct its temporal evolution and predict its future states. We validate our approach on diverse spheres of human activities ranging from scientific and artistic productivity, to sexual relations and online traffic.

We then broaden our framework to not only focus on resource distribution, but to also consider the resulting structure. We thus apply our framework to the theory of complex networks which describes the connectivity structure of social, technological or biological systems. In so doing, we propose that an important class of complex systems can be modelled as a construction of potentially infinitely many levels of organization all following the same universal growth principle known as preferential attachment. We show how real complex networks can be interpreted as a projection of our model, from which naturally emerge not only their scale independence, but also their clustering or modularity, their hierarchy, their fractality and their navigability. Our results suggest that social networks can be quite simple, and that the apparent complexity of their structure is largely a reflection of the complex hierarchical nature of our world.

Contents

List of Tables

List of Figures

À qui de droit.

I have no idea where this will lead us, but I have a definite feeling it will be a place both wonderful and strange.

Dale Cooper, Twin Peaks.

Remerciements

Je me dois tout d'abord de remercier ceux qui ont rendu cette thèse possible: le groupe Dynamica. En premier lieu, notre directeur Louis J. Dubé sans qui nous n'aurions ni l'environnement, ni le support, ni la liberté d'aborder des problèmes aussi variés et parfois risqués. En second lieu, ceux avec qui les travaux ont été effectués, la section réseaux: Antoine Allard, Jean-Gabriel Young, Vincent Marceau, Pierre-André Noël et Edward Laurence. Leur nom devrait figurer au côté du mien sur la page titre de cette thèse. En dernier lieu, le reste du groupe, tout aussi important: Denis Gagnon, Guillaume Painchaud-April, Joey Dumont et Jean-Luc Déziel.

Je me dois aussi de remercier tous ceux qui m'ont apporté support pendant les dernières 5 années. Pour le financement, l'IRSC, le FQRNT et le CRSNG. Pour le support moral et quotidien, ma mère Danielle Hébert, mon père Guy Dufresne, ma soeur Lysiane Hébert-Dufresne et mes chums de boisson. Je dois finalement une dernière mention à Juniper Lovato pour l'inspiration, la motivation et l'aide apportées pendant l'écriture.

Avant-propos

Un peu tout comme la science de la complexité elle-même, le travail effectué pendant mes études graduées transcende les carcans de la physique plus traditionnelle: théorie des réseaux complexes, distribution de ressources dans des systèmes sociaux et propagation de maladies. Cette thèse n'en couvrira malheureusement qu'une partie, mais le choix a été fait afin de raconter une histoire cohérente et complète. Cela étant dit, les chapitres peuvent pour la plupart être lus de façon indépendante, ce qui implique que certains concepts de base sont répétés et revisités abondamment. De même, les annexes offrent soient des discussions supplémentaires ou des détails sur les méthodes utilisées, elles sont donc présentées par souci d'exhaustivité, mais ne devraient nullement être nécessaires à la lecture de la thèse principale. Notez également que, après ce court avant-propos, le texte est entièrement en anglais à l'exception de résumés introduisant chaque chapitre.

La thèse couvrira les travaux effectués sur les processus de croissance de systèmes indépendants d'échelle ainsi que leurs liens avec les distributions de ressources sociales et avec la théorie des réseaux complexes. Le *premier chapitre* aborde l'*indépendance d'échelle* de façon très générale. Nous présentons également certaines notions théoriques qui s'y rattachent, illustrées à l'aide de quelques exemples empiriques.

Le *second chapitre* est une courte revue de la littérature sur l'*attachement préférentiel*: un processus de croissance décrivant l'évolution d'un système vers l'indépendance d'échelle en termes d'un jeu de balles et d'urnes. Ce processus découle des éléments théoriques présentés dans le chapitre précédent, et est au coeur de cette thèse. Notons au passage que la section 2.2.1 est une version légèrement modifiée d'une section d'une de nos publications [57].

Le *troisième chapitre* aborde l'évolution d'une distribution vers un état indépendant d'échelle. Nous démontrons que cette évolution implique certaines contraintes temporelles. À l'exception des annexes, ce chapitre est une version préliminaire d'une future publication [60]. Le *quatrième chapitre* complète cette portion sur la *complexité temporelle* de l'attachement préférentiel en présentant comment notre théorie inclut divers modèles classiques. Nous discutons également du danger d'étudier les détails microscopiques de systèmes complexes en se basant sur leurs propriétés macroscopiques. Les résultats des Sec. 3.6 et 4.3 ont été préalablement présentés en conférence [55].

Les deux derniers chapitres concernent la *complexité structurelle* de l'attachement préférentiel. Le *cinquième chapitre* montre comment les travaux précédents peuvent être généralisés pour décrire certaines propriétés non triviales de la structure des réseaux sociaux et technologiques. Les sections 5.1 et 5.2 sont tirés d'une lettre, et de son matériel supplémentaire, publiée en 2011 [56] dans le cadre de mes travaux de maîtrise. Dans le *sixième et dernier chapitre*, nous généralisons encore d'avantage nos travaux pour illustrer l'importance du concept de *hiérarchie*. En considérant maintenant un attachement préférentiel hiérarchique, nous réduisons significativement la complexité apparente des propriétés structurelles des réseaux complexes. L'essentiel de ce chapitre est une version préliminaire d'une future publication [61].

Finalement, la liste suivante contient l'ensemble des publications auxquelles j'ai participé pendant mes études graduées. Elles sont présentées en ordre chronologique inverse. Toutes les publications insérées dans cette thèse sont marquées d'un astérisque. Dans tous les cas, j'ai agit en tant que premier auteur. C'est-à-dire que j'ai été responsable de la conception du projet, de la modélisation mathématique, de la programmation, de la production de résultats et de la rédaction du texte avec l'aide de mes co-auteurs. La seule exception étant la programmation du modèle d'attachement préférentiel hiérarchique effectuée par Edward Laurence. Les autres auteurs sont ordonnés selon l'importance de leur contribution, à l'exception de notre directeur de recherche qui est présenté comme dernier auteur dans tous les cas.

24. **L. Hébert-Dufresne**, V. Marceau, P.-A. Noël, A. Allard & L.J. Dubé
The social zombie: Modelling undead outbreaks on social networks.
Mathematical Modelling of Zombies, Ed. Robert Smith?, University of Ottawa Press (2014).

23. **L. Hébert-Dufresne**, J.-G. Young, A. Allard & L.J. Dubé
Structural preferential attachment of community structure and its relation to Dunbar's number*.
en préparation.

22. **L. Hébert-Dufresne**, E. Laurence, A. Allard, J.-G. Young & L.J. Dubé
Complex networks are an emerging property of hierarchical preferential attachment*.
arXiv, en soumission.

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

20. **L. Hébert-Dufresne**, A. Allard, J.-G. Young & L.J. Dubé
Universal growth constraints of human systems*.
arXiv, en soumission.

19. G.M. Goerg, O. Patterson-Lomba, **L. Hébert-Dufresne** & B.M. Althouse
Escaping the poverty trap: modeling the interplay between economic growth and the ecology of infectious disease.
arXiv, en soumission.

18. J.-G. Young, A. Allard, **L. Hébert-Dufresne** & L.J. Dubé
Unveiling Hidden Communities Through Cascading Detection on Network Structures.
arXiv, en soumission.

17. B.M. Althouse & **L. Hébert-Dufresne**
Epidemic Cycles Driven by Host Behavior.
Journal of the Royal Society Interface 11 20140575 (2014).

16. A. Allard, **L. Hébert-Dufresne**, J.-G. Young & L.J. Dubé
Coexistence of phases and the observability of random graphs.
Physical Review E 89 022801 (2014).

15. P.-A. Noël, A. Allard, **L. Hébert-Dufresne**, V. Marceau, & L.J. Dubé
Spreading dynamics on complex networks: a general stochastic approach.
Journal of Mathematical Biology (2013)

14. **L. Hébert-Dufresne**, A. Allard, J.-G. Young & L.J. Dubé
Percolation on random networks with arbitrary k-core structure.
Physical Review E 88 062820 (2013).

13. **L. Hébert-Dufresne**, A. Allard, J.-G. Young & L.J. Dubé
Global efficiency of local immunization on complex networks.
Scientific Reports, 3, 2171 (2013).

12. **L. Hébert-Dufresne**, O. Patterson-Lomba, G.M. Goerg & B.M. Althouse
Pathogen mutation modeled by competition between site and bond percolation.
Physical Review Letters, 110, 108103 (2013).

11. B.M. Althouse, O. Patterson-Lomba, G.M. Goerg & **L. Hébert-Dufresne**
Targeting and timing of treatment influences the emergence of influenza resistance in structured populations.
PLOS Computational Biology, 9(2), e1002912 (2013).

10. O. Patterson-Lomba, B.M. Althouse, G.M. Goerg & **L. Hébert-Dufresne**
Optimizing treatment regimes to hinder antiviral resistance in influenza across time scales.
PLOS ONE 8(3): e59529 (2013).

9. **L. Hébert-Dufresne**, A. Allard, J.-G. Young & L.J. Dubé
On the constrained growth of complex critical systems*.
arXiv, préparé pour la 2nd International Conference on Complex Sciences: Theory and Applications (Santa Fe, NM, 2012).

8. **L. Hébert-Dufresne**, A. Allard, V. Marceau, P.-A. Noël & L.J. Dubé
Stochastic process for the growth of scale-free, modular and self-similar systems*.
Physical Review E, 85, 026108 (2012).

7. A. Allard, **L. Hébert-Dufresne**, P.-A. Noël, V. Marceau & L.J. Dubé
Bond percolation on a class of correlated and clustered random graphs.
Journal of Physics A: Mathematical and Theoretical 45, 405005 (2012).

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

5. P.-A. Noël, A. Allard, **L. Hébert-Dufresne**, V. Marceau & L.J. Dubé
   Propagation on networks: an exact alternative perspective.
   Physical Review E, 85, 031118 (2012).

4. **L. Hébert-Dufresne**, A. Allard, V. Marceau, P.-A. Noël & L.J. Dubé
   Structural preferential attachment: Network organization beyond the link*.
   Physical Review Letters, 107, 158702 (2011).

3. V. Marceau, P.-A. Noël, **L. Hébert-Dufresne**, A. Allard & L.J. Dubé
   Modeling the dynamical interaction between epidemics on overlay networks.
   Physical Review E, 84, 026105 (2011).

2. V. Marceau, P.-A. Noël, **L. Hébert-Dufresne**, A. Allard & L.J. Dubé
   Adaptive networks: coevolution of disease and topology.
   Physical Review E, 82, 036116 (2010).

1. **L. Hébert-Dufresne**, P.-A. Noël, V. Marceau, A. Allard & L.J. Dubé
   Propagation dynamics on networks featuring complex topologies.
   Physical Review E, 82, 036115 (2010).

Foreword

As complexity science itself, the work done during my years in graduate school attempts to transcend the usual boundaries of physics: complex network theory, distribution of resources in social systems and disease propagation among others. This thesis will unfortunately only cover parts of it, but the choice was made in order to tell a more complete and coherent story. The chapters can mostly be read independently, which also implies that most basic concepts will be repeated and revisited abundantly. The appendices might offer additional discussions or details on our methods — and are thus presented for the sake of completeness — but should not be necessary to the reading of our main thesis.

More precisely, the thesis will cover our work on stochastic growth processes of scale independent systems as well as their link with the distributions of social resources and the theory of complex networks. The *first chapter* tackles scale independence in a very general way. It also covers related theoretical notions, illustrated with empirical examples.

The *second chapter* is a short literature review on preferential attachment: a growth process describing the evolution of a system towards scale independence, in terms of a balls in urns game. This process follows logically from the theoretical elements presented in the preceding chapter, and is at the very heart of this thesis. Note that section 2.2.1 is a slightly modified version of a section previously published by our group [57].

The *third chapter* deals with the evolution of a distribution towards scale independence. We show how this evolution implies certain strict constraints. With the exception of the appendices, this chapter consists of a preliminary version of a future publication [60]. The *fourth chapter* complements this portion on the *temporal complexity* of preferential attachment by presenting how our framework includes diverse classical models. We also discuss the danger of using macroscopic properties of complex systems to infer some microscopic details. Results of Sec. 3.6 and 4.3 were previously presented in conference [55].

The last two chapters concern the *structural complexity* of preferential attachment. The *fifth chapter* illustrates how to generalize our previous efforts to now describe non-trivial properties of social and technological networks. Sections 5.1 and 5.2 are a reproduction of a letter, and of its supplementary material, published in 2011 [56] as part of my master's thesis. In the *sixth and final chapter*, we generalize our work further to highlight the importance of *hierarchy*. By now considering a hierarchical preferential attachment, we significantly reduce the apparent complexity of the structural properties observed in complex networks. The essential of this chapter is a preliminary version of a future publication [61].

Finally, the following list contains all scientific publications produced during my postgraduate studies. It is presented in inverse chronological order. All publications included in this thesis are marked with an asterisk. In all cases, I acted as first author and was responsible of the project conception, mathematical modelling, programming, results production and of the redaction of the manuscript with help from my co-authors. The only exception being the programming of the hierarchical preferential attachment model by Edward Laurence. The other authors are always listed according to the importance of their contributions, with the exception of our research advisor who acted as senior author in all cases.

24. **L. Hébert-Dufresne**, V. Marceau, P.-A. Noël, A. Allard & L.J. Dubé
The social zombie: Modelling undead outbreaks on social networks.
Mathematical Modelling of Zombies, Ed. Robert Smith?, University of Ottawa Press (2014).
23. **L. Hébert-Dufresne**, J.-G. Young, A. Allard & L.J. Dubé
Structural preferential attachment of community structure and its relation to Dunbar's number*.
in preparation.
22. **L. Hébert-Dufresne**, E. Laurence, A. Allard, J.-G. Young & L.J. Dubé
Complex networks are an emerging property of hierarchical preferential attachment*.
arXiv pre-print, in submission.
21. A. Allard, **L. Hébert-Dufresne**, J.-G. Young & L.J. Dubé
A general and exact approach to percolation on random graphs.
in submission.
20. **L. Hébert-Dufresne**, A. Allard, J.-G. Young & L.J. Dubé
Universal growth constraints of human systems*.
arXiv pre-print, in submission.
19. G.M. Goerg, O. Patterson-Lomba, **L. Hébert-Dufresne** & B.M. Althouse
Escaping the poverty trap: modeling the interplay between economic growth and the ecology of infectious disease.
arXiv pre-print, in submission.
18. J.-G. Young, A. Allard, **L. Hébert-Dufresne** & L.J. Dubé
Unveiling Hidden Communities Through Cascading Detection on Network Structures.
arXiv pre-print, in submission.
17. B.M. Althouse & **L. Hébert-Dufresne**
Epidemic Cycles Driven by Host Behavior.
Journal of the Royal Society Interface 11 20140575 (2014).
16. A. Allard, **L. Hébert-Dufresne**, J.-G. Young & L.J. Dubé
Coexistence of phases and the observability of random graphs.
Physical Review E 89 022801 (2014).

15. P.-A. Noël, A. Allard, **L. Hébert-Dufresne**, V. Marceau, & L.J. Dubé
Spreading dynamics on complex networks: a general stochastic approach.
Journal of Mathematical Biology (2013)

14. **L. Hébert-Dufresne**, A. Allard, J.-G. Young & L.J. Dubé
Percolation on random networks with arbitrary k-core structure.
Physical Review E 88 062820 (2013).

13. **L. Hébert-Dufresne**, A. Allard, J.-G. Young & L.J. Dubé
Global efficiency of local immunization on complex networks.
Scientific Reports, 3, 2171 (2013).

12. **L. Hébert-Dufresne**, O. Patterson-Lomba, G.M. Goerg & B.M. Althouse
Pathogen mutation modeled by competition between site and bond percolation.
Physical Review Letters, 110, 108103 (2013).

11. B.M. Althouse, O. Patterson-Lomba, G.M. Goerg & **L. Hébert-Dufresne**
Targeting and timing of treatment influences the emergence of influenza resistance in structured populations.
PLOS Computational Biology, 9(2), e1002912 (2013).

10. O. Patterson-Lomba, B.M. Althouse, G.M. Goerg & **L. Hébert-Dufresne**
Optimizing treatment regimes to hinder antiviral resistance in influenza across time scales.
PLOS ONE 8(3): e59529 (2013).

9. **L. Hébert-Dufresne**, A. Allard, J.-G. Young & L.J. Dubé
On the constrained growth of complex critical systems*.
arXiv pre-print, prepared for the 2nd International Conference on Complex Sciences: Theory and Applications (Santa Fe, NM, 2012).

8. **L. Hébert-Dufresne**, A. Allard, V. Marceau, P.-A. Noël & L.J. Dubé
Stochastic process for the growth of scale-free, modular and self-similar systems*.
Physical Review E, 85, 026108 (2012).

7. A. Allard, **L. Hébert-Dufresne**, P.-A. Noël, V. Marceau & L.J. Dubé
Bond percolation on a class of correlated and clustered random graphs.
Journal of Physics A: Mathematical and Theoretical 45, 405005 (2012).

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

5. P.-A. Noël, A. Allard, **L. Hébert-Dufresne**, V. Marceau & L.J. Dubé
Propagation on networks: an exact alternative perspective.
Physical Review E, 85, 031118 (2012).

4. **L. Hébert-Dufresne**, A. Allard, V. Marceau, P.-A. Noël & L.J. Dubé
Structural preferential attachment: Network organization beyond the link*.
Physical Review Letters, 107, 158702 (2011).

3. V. Marceau, P.-A. Noël, **L. Hébert-Dufresne**, A. Allard & L.J. Dubé
Modeling the dynamical interaction between epidemics on overlay networks.
Physical Review E, 84, 026105 (2011).

2. V. Marceau, P.-A. Noël, **L. Hébert-Dufresne**, A. Allard & L.J. Dubé
Adaptive networks: coevolution of disease and topology.
Physical Review E, 82, 036116 (2010).

1. **L. Hébert-Dufresne**, P.-A. Noël, V. Marceau, A. Allard & L.J. Dubé
Propagation dynamics on networks featuring complex topologies.
Physical Review E, 82, 036115 (2010).

Introduction

This thesis tackles subjects that could easily be tagged with plenty of currently popular buzzwords: complexity, emergent behaviours, universal properties or self-organized criticality, to name a few. In fact, I will definitely not shy away from using such phrases; but I also feel that they should come with a word of warning.

The science of complexity can be aptly described as the science of emergent behaviours; which are in return often described through the motto that *the whole is more than the sum of its parts*. Studying the distribution of resources and activities in social systems certainly fits the bill. Describing such dynamics from the bottom up — perhaps based on our knowledge of individual behaviour, on psychological principles or on our limited knowledge of neuropsychology — would be a challenge of Herculean proportion. Yet, that is more a problem of perspective than a limit of the reductionist approach.

There is no arguing that complexity science does highlight many of the conceptual and theoretical limits of reductionism. For instance in network theory, where the properties of a system stem not from its elements but from the structure of their interconnections. Or in epidemic dynamics, where evolution of pathogens and human behaviours co-evolve and adapt. Or in any coupling of fundamentally different systems such as the interactions of economics with our technological and biological environment. Those are fields and problems where holism has led to important advances. Yet, there remains a danger in dismissing out-of-hand a philosophy which has lead science to so many successes. Once again, the challenge is mostly in selecting the right perspective.

In fact, this thesis follows a long line of effort in using statistical physics to describe social systems independently of the social nature of the interactions in very much the same way it allows the development of thermodynamics without describing interactions at the molecular level. While this is a fitting analogy, we rarely see thermodynamics ranked alongside network theory, chaos, artificial intelligence or any other flagship of complexity science. Yet, even to the most brilliant observer, phase transitions and the second law of thermodynamics are surprising emergent behaviours. While these properties are far from intuitive consequences of how gases mix and colliding particles exchange energy, they can still be described under a more appropriate perspective: following the distribution of energy to gather *macroscopic* properties without the corresponding *microscopic* description. In many ways, this is exactly the approach that we apply to social systems.

Statistical physics of social systems

Statistical physics deals with ensembles of populations. That is, its goal is to reach a statistical description of how a population evolves in a given problem. The population itself is usually defined through a set of statistical properties, while the problem under study is defined through constraints on the statistical solutions. Once both are given, statistical physics aims to describe the ensemble of all possible scenarios for a given population.

To draw upon physical examples, populations often refer to a set of particles whose statistical properties describe how they should be counted and how they can be organized. Are they distinguishable or indistinguishable? Are they classical or quantum in nature? How many states can they occupy and under what intrinsic constraints? For instance, a quantum description of an electron population in a solid would consider an indistinguishable population occupying a discrete set of states under the Pauli exclusion principle (it is impossible for two electrons to occupy the exact same state). We then know that the ensemble will be described through Fermi-Dirac statistics. In the context of a statistical physics of social systems, our population will almost always be defined as follows. Individuals (particles) share discrete amounts of activity or resources (discrete energy states) and individuals with equal shares are indistinguishable.

While these properties inform our choice of statistical methods, the exact solution of the ensemble depends on the constraints imposed by the physical problem. To name only the quintessential case: the canonical ensemble represents systems in contact with a heat bath, thus fixing the average energy. The exact solution for the distribution of energy in the population can then be obtained by maximizing entropy while respecting both normalization and the fixed average energy (zero-th and first moments). Following previous work, our statistical ensemble of social systems is also defined through a maximization of entropy, under finite size constraints, and considering our particular population definition.

The first chapter builds the ensemble on which we will work. The goal here is to develop a conceptual understanding of why scale-independent systems, social or otherwise, deserve to be studied. We also illustrate how statistical physics appears to be a good tool for the task and we discuss some of the mathematical subtleties involved without lingering too much in details.

Interplay of growth and organization

At this point, it is interesting to note that this "statistical physics of scale-independent systems" follows the development of classical statistical physics almost in reverse. Applications of heat transfer came first, without any understanding of thermodynamics as we know it. Only much later did we develop the statistical physics framework and we had to wait even further before observation of its predicted universality [1]. For scale-independent systems, empirical observations of its universal steady state came first. We now have theories to explain the existence of that steady state, but still without an understanding of the underlying physical processes.

To this end, the second chapter presents previous work on the so-called preferential attachment principle: a simple growth process that gives rise to scale independence. This growth process is attractive within our framework for multiple reasons. First, it is incredibly simple. It relies on the same assumption as the definition of our population, namely that only individuals with different shares can be distinguished from one another. Second, it is analytically tractable through standard tools of statistical physics; and we will here rely heavily on the rate equation (or master equation) approach. Third, it can be shown to not only be a sufficient cause for scale independence, but that under our basic assumptions, scale independence necessarily implies a preferential attachment process.

This last point is presented in the third chapter, which describes how a power-law organization can constrain a system's growth. This chapter and the following are in fact devoted to our main contribution: the interplay between the organizational scale independence of a system and the temporal properties of its growth.

A new point of view for complex networks

The last two chapters investigate the structure of the previously considered activities. Consider one of the systems that we will often revisit: sexual encounters between clients and escorts. We can simply focus on the distribution of sexual activity and its growth, or we can actually delve into the system and study the structure. Who had sexual relations with whom? How can we use our previous framework to start describing what this sexual contact network, or any other social networks, might actually look like?

We tackle this question in a very loose but hopefully insightful manner. Simply put, we attempt to answer the following question: why use networks? Why is so much emphasis given to the actual links between nodes? Network scientists sometimes have a strange modus

[1] *Universality* will here always refer to properties that appear independent of the system's nature or microscopic details.

operandi. First, collapse a system to a network, basically projecting all structures, correlations, and social hierarchy as a set of lines between nodes; then try to infer structure from this set of lines. Here we show how multiple emergent properties of real networks, which are essentially why complex networks are called complex, can be understood as artefacts of the hierarchical nature of our world. To illustrate this, we generalize our previous framework to show how embedded levels of activities (say, a researcher involved in universities, research groups, and individual papers) give rise to complex structural properties. This suggests that obtaining a crude description of a system's hierarchy is perhaps a better way to describe its underlying network than directly looking at the links between its elements.

While these chapters are in a lot of ways a proof of concept, they certainly widen the applicability and implications of the work presented in the first four chapters.

Final word before we begin

This thesis is far from a definitive description of the statistical physics of scale-independent social systems. Even if or why these systems might be scale-independent at all is arguable. Our contributions are in no way dependent on *if* or *why* a given system is scale-independent, but only on the fact that a power-law distribution is a good enough approximation of their organization at some level. Unfortunately, how good is "good enough" remains to be determined; we are mostly gaining insights on how highly heterogeneous organizations influence diverse temporal and structural properties. The extent to which these insights can be applied outside of the context where they were derived is at times surprising and always of great use.

Chapter 1

On scale independence

Résumé

Cette thèse est entièrement dédiée à l'étude des systèmes qui s'auto-organisent de façon *critique* ou *indépendante d'échelle*. Dans ce contexte, la criticalité et l'indépendance d'échelle font toutes deux référence à l'absence d'échelle caractéristique. Ce chapitre définit donc les concepts reliés à cette propriété et les illustre avec des systèmes particuliers. Par exemple, nous illustrons comment l'indépendance d'échelle est liée aux transitions de phases à l'aide d'une courte introduction à la théorie de la percolation.

Nous présentons également les outils statistiques qui seront utilisés par la suite: estimation du maximum de vraisemblance et distance statistique entre distributions de probabilité. Ceux-ci sont tirés des travaux de Clauset, Shalizi et Newman [28]. Nous abordons aussi certains éléments théoriques qui serviront de tremplin à nos travaux, dont la théorie de Baek, Bernhardsson et Minnhagen sur l'universalité de l'indépendance d'échelle [8]. En somme, les objectifs de ce chapitre sont de définir ce qu'est l'indépendance d'échelle, de développer une intuition qualitative de ce qu'elle implique, d'être en mesure de la quantifier, et d'expliquer pourquoi cette propriété est intéressante et si répandue.

Summary

This entire thesis is dedicated to systems that self-organize into a *scale-independent*, *scale-free* or *critical* organization. Here, these three concepts all refer to essentially the same idea: the lack of characteristic scale. In this chapter, we define these terms and illustrate them by drawing upon empirical examples. For instance, we qualitatively illustrate the connection between scale independence and phase transitions through a basic introduction of percolation theory.

We also follow the work of Clauset, Shalizi and Newman [28] to develop the statistical tools required for the rest of our work: maximum-likelihood estimation and statistical distance between distributions. Moreover, we present elements of theory which will serve as the foundation of our work, such as the theory of Baek, Bernhardsson and Minnhagen on the universality of scale independence [8]. The objectives of this chapter are thus to define scale independence, to develop a qualitative intuition of its implications, to quantify it and to explain why it is so ubiquitous and worthy of interest.

1.1 Scale independence, power laws, and criticality

1.1.1 Scale-independent organization

Strictly speaking, scale independence refers to the property of a real function $f(x)$ (for $x \in \mathbb{R}$) to repeat itself at any scale. This means that when there is scale independence, a dilatation or contraction of the scale of interest, for instance going from $x \in [1, 10]$ to $[100, 1000]$, only implies a rescaling of the observed function f. A function can thus be defined as scale-independent if it obeys the property

$$f(\lambda x) = \lambda^{\gamma} f(x) \ . \tag{1.1}$$

This behaviour is interesting for multiple reasons. First and foremost, because it is ubiquitous amongst natural self-organized systems [78]. Second, its mathematical description is in itself also interesting for its simplicity. In fact, a single class of differentiable functions features scale independence, *power laws*, i.e., functions of the form

$$f(x) \propto ax^{\gamma} \ . \tag{1.2}$$

When considering the distribution $\{p_k\}$ of some random variable k, as will be done throughout the thesis, a pure power-law distribution is essentially solely defined by its parameter γ. We will typically refer to γ as the scale exponent of the power law since it sets the form of the rescaling under which Eq. (1.1) is respected. The other parameter, a, is then simply set as a normalization factor to assure that the distribution summed over all k equals one. For instance, in the case of a discrete quantity k, we would write

$$p_k = \frac{k^{-\gamma}}{\sum_{k>0} k^{-\gamma}} \tag{1.3}$$

where the minus sign in the exponent is introduced to explicitly show that the distribution must be decreasing for the sum to converge. In fact, the interest for the scale exponent γ lies in how it clearly discriminates which moments of the distribution will converge and which will diverge to infinity. For instance, in the case of a continuous random variable x with probability density $p(x) = ax^{-\gamma}$, we would set the normalization factor through

$$1 = \int_1^\infty ax^{-\gamma}dx = \frac{a}{1-\gamma}\left[x^{1-\gamma}\right]_1^\infty \ . \tag{1.4}$$

The right hand side of the last expression is finite if and only if $\gamma > 1$, which ensures that $x^{1-\gamma}$ vanishes. The normalization condition is then satisfied by setting

$$a = \gamma - 1 \ \Rightarrow p(x)dx = ax^{-\gamma}dx. \tag{1.5}$$

Note that we use $x = 1$ as the lower bound for our variable as this will be the case in the discrete distributions that will be hereafter studied. However, we could have just as easily used any arbitrary bound $x_{\text{min}} > 0$ to avoid the divergence at $x = 0$. In a similar fashion, the expected mean value $\langle x \rangle$ would be obtained from

$$\langle x \rangle = \int_1^\infty xp(x)dx = a\int_1^\infty x^{1-\gamma}dx = \frac{\gamma-1}{\gamma-2} \qquad \text{for } \gamma > 1 \tag{1.6}$$

which this time is finite if $\gamma > 2$. The same logic applies to all other moments,

$$\langle x^m \rangle = \int_1^\infty x^m p(x)dx = a\int_1^\infty x^{m-\gamma}dx = \frac{\gamma-1}{\gamma-m-1} \qquad \text{for } \gamma > m+1 \ . \tag{1.7}$$

This means that the m-th moment of a power-law distribution always exists if and only if its scale exponent is greater than $m + 1$. This result holds also for discrete distributions. This is straightforwardly shown as the divergence of the integral (or sum) always occurs at $x \to \infty$ (or $k \to \infty$). Since power laws are slowly varying functions, one can always evaluate a discrete sum exactly for all values $k < k^*$, with an appropriate bound $k^* \gg 1$, then approximate the remaining portion as an integral on the interval $[k^*, \infty[$. As the previous logic is independent of x_{min}, the result on the convergence of the m-th moment must hold also for discrete distributions.

1.1.2 Scale-free organization of scale-independent systems

Beyond the elegance of rescaling, we mentioned in the introduction that scale independence also refers to the lack of a characteristic scale. Hence, these systems are often referred to as scale-free. To place this in perspective, we can juxtapose scale-independent distributions with more common distributions [96]. Consider, for instance, the two following questions.

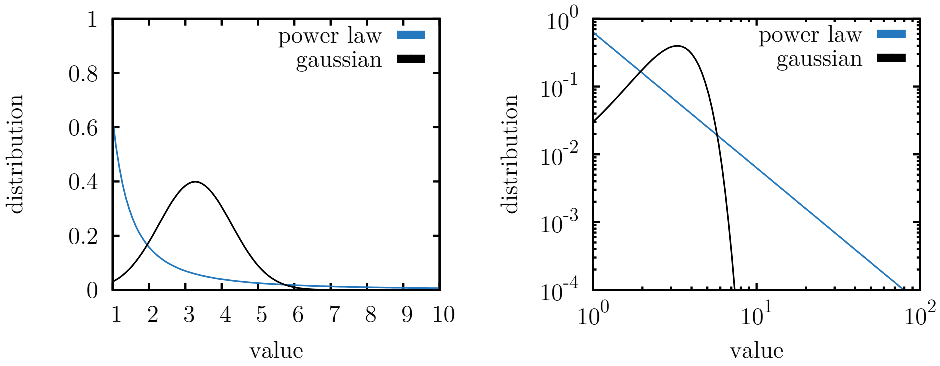


Figure 1.1 – **Power-law and normal distributions.** Visual comparisons of a power-law distribution of scale exponent $\gamma = 2$ with a normal distribution: (left) linear scales, (right) logarithmic scales. The log-log plot is often used to represent power laws visually as $y = x^\gamma$ becomes $\log(y) = \gamma\log(x)$: a simple straight line whose slope is given by the scale exponent.

- What is the probability that a randomly selected adult is twice as tall as the average adult? Most likely zero. In fact, the probability for anyone to measure ten times the average adult height is *exactly* zero.
- What is the probability that a randomly selected adult is twice as wealthy as the average adult? Hard to say, but definitely non-zero. In fact, there is a sizeable number of people whose fortune will be at least ten, or even a thousand times larger than the average.

The second question concerns the type of behaviour that Benoît Mandelbrot called wild or extreme randomness [72] [1]. The distribution of adult human height is well-behaved, with a finite mean, variance, and higher moments. In fact, while there is significant variance between sex or ethnic groups, height distribution on a given population roughly follows a normal distribution. On the other hand, wealth distribution roughly follows a power law, which in this context is sometimes called a Pareto distribution. In this particular case, the organization of wealth has a scale exponent greater than 1 but less than 2, meaning that in the limit of an infinite population it would have an infinite mean. In other words, following a given dollar used in a given transaction, that dollar would, on average, end up in the wallet of an infinitely wealthy person! Thus, this person is at least ten times richer than the average person. Figure 1.1 visually compares an example of a power-law distribution with the more familiar normal distribution.

The *Pareto law*, the power-law-like distribution of wealth, is the cause of the well-known 80-20 principle, or Pareto principle, which is where the law takes its name [85]. Pareto, an Italian economist, originally made the observation that around 80% of Italian land in 1906 was owned by around 20% of the population. Similarly, he noted, 20% of pea pods in his garden contained 80% of peas. This led him to formulate his principle for the division of

[1] More precisely, Mandelbrot defined “wild randomness” as events with diverging second moments but finite mean, and “extreme randomness” for distribution where all moments are diverging.

a resource in a population. In fact, the extreme nature of his observations stems from the underlying power-law distribution, where the 80-20 division is merely a consequence of their particular scale exponent. A generalized version of this principle could state something like *a microscopic fraction of the population shares a macroscopic fraction of the resource.* In that case, resource implies that we are summing over k (i.e., the first moment $\sum kp_k$), but the same principle applies just as well to higher moments. In fact, the higher the moment, the greater the inequality.

What does this mean for other power-law distributed quantities? Earthquakes are a good example of this phenomenon in the natural world [96]. Without delving into the debate over its exact form and interpretation, we can safely state that the tail of the distribution of magnitude per earthquakes roughly follows a power law with a scale exponent around 5/3 [53]. Using our previous results for $\gamma < 2$, this implies that the average earthquake has an infinite magnitude! Finite available energy notwithstanding, this does not mean that a finite sequence of earthquakes would have an infinite magnitude, but simply that in a potentially infinite sequence, the sum of magnitudes would grow faster than the number of earthquakes. We can play with this subtlety using simple statistics [2] and try to tame the wild randomness of earthquakes.

Using our previous results, we can say that the (continuous) distribution of earthquake magnitude follows $P(x) = 2x^{-5/3}/3$ using $x_{\text{min}} = 1$ as even those are usually not recorded by seismologists. We can thus write its cumulative distribution function $C(x)$ as

$$C(x) = \int_1^x P(x')dx' = \frac{2}{3}\int_1^x (x')^{-5/3}dx' = 1 - x^{-2/3} \ . \tag{1.8}$$

In a discrete distribution, a finite but very large sequence of N events, can be expected to feature roughly $N\sum_{k'=k-\Delta}^{k+\Delta} p_{k'}$ events falling within an interval 2Δ around event k. In the continuous case, with an interval $\Delta = x$, we would expect $NC(2x)$ events of magnitude less than $2x$. Similarly, we would be very surprised to see magnitudes above whatever x_c marks $NC(x_c) = N - 1$, expecting on average only one event above that value. With that logic, solving for x_c

$$C(x_c) = 1 - x_c^{-2/3} = 1 - \frac{1}{N} \quad \rightarrow x_c = N^{3/2} \ , \tag{1.9}$$

yields an idea of the strongest expected earthquake. We can now roughly estimate the average magnitude in a finite sequence of N earthquakes:

$$\langle x \rangle_N = \int_1^{x_c} x'P(x')dx' = \int_1^{x_c} \frac{2}{3}(x')^{-2/3}dx' \simeq 2x_c^{1/3} = 2\sqrt{N} \tag{1.10}$$

which tells us that, as we sample the distribution, the mean is expected to go to infinity as fast as a square root! The most peculiar result we can obtain from this analysis is the expected fraction of the total energy unleashed that was contained in the single largest earthquake:

$$\left\langle \frac{x_c}{\sum_i x_i} \right\rangle \approx \frac{x_c}{N\langle x \rangle_N} = \frac{N^{3/2}}{2N\sqrt{N}} = \frac{1}{2} \ . \tag{1.11}$$

[2] . . . and assuming that earthquakes are independent and identically distributed.

This is just an order of magnitude as additional care must be taken when averaging. Indeed, x_c is not independent of the sum in the denominator. Yet, the main point is clear: no matter how large our "infinite" sequence is, the largest earthquake therein holds a finite fraction of all the energy!

This small exercise was only meant to emphasize the intricacies of power-law analysis. The mode of the magnitude distribution (i.e., the most probable event) is at the lowest considered magnitude and we can therefore expect a huge concentration of earthquakes at that point. Yet, we can also expect a single event several orders of magnitude bigger (in this case $\sim N^{3/2}$) containing as much energy as the sum of all the smaller earthquakes.

This is why the power-law organization is *scale-free*. All possible scales must be considered; there is no characteristic scale. More conceptually, this also implies that the power-law distribution is a very critical organization in-between a highly homogeneous state — most events around $x_{\min}$ — and a highly heterogeneous state — one event containing most of the energy.

1.1.3 Criticality in phase transitions and percolation

The interpretation of scale independence as an organization in-between states is nothing new. It is actually at the very core of *phase transition theory*. This theory is a formal study of the transformation of a system from one state to another. In our everyday life, we have some experience of phase transitions, e.g when ice melts (solid → liquid) or when steam condenses on a window (gas → liquid). Unfortunately, these experiences do not necessarily help in determining what occurs at the exact point of the transition.

In this section, we will study this question in a very conceptual way. The goal here is to provide insights into why a scale-free organization is *in-between states*. To illustrate our point, we will focus on a two-state system which is either disorganized (fluid) or organized (solid), and model its transition using basic percolation theory.

Percolation theory is a stochastic process of diffusion in a disordered media [26]. Unlike actual diffusion and Brownian motion, the randomness does not reside in the moving particles/fluid, but in the medium that holds them. Percolation can thus be used as a toy model for anything propagating with trivial rules within a non-trivial structure, even forest fires [38] and epidemics [77, 73, 7, 63, 58]. As we will see, the problem is extremely simple to state, but usually not exactly solvable.

To quickly define percolation, it is perhaps useful to fall back on its most familiar expression: the brewing of coffee [39]. Imagine space as a three dimensional grid or mesh, where every site is either occupied by a particle of coffee or potentially by a drop of water. Since in this case the diffusing fluid is the water and not the coffee, we will call *occupied* a site that holds water. Each site is occupied[3] independently with a given probability u, such that any occupied site can be isolated or form groups (*clusters*) with its nearest-occupied-neighbours.

If $u \ll 1$, most clusters will be small and isolated, whereas if $u \sim 1$ most sites will form a giant *spanning cluster* extending from one end of the grid to the other. For an infinite medium, the size of this spanning cluster will be measured by the fraction of all sites which are part of this giant cluster. This fraction will thus be our order parameter. If non-zero, we are in an ordered phase where a path exist across the medium and we get coffee! The spanning cluster marks the way water can diffuse through the coffee and reach the other end (our cup). If zero, no path exists and we find ourselves in a disordered phase where all clusters are finite and isolated. In that case, all we obtain is a wet coffee sludge. In fact, a very precise critical value u_c marks the phase transition. If $u \geq u_c$, a spanning cluster exists and we end up in the ordered state.

During the production of this thesis, we investigated several variations on this basic percolation process: solving the process on arbitrary finite structure [5], infinite general structure [4] or by modifying the percolative process itself [63, 6]. While the structure and process may change, the interest is always in the behaviour at the percolation threshold. As a simple example, we will here study the case of percolation in one dimension which can be solved analytically [39].

In one dimension, it is easy to see how the probability for one site to be occupied is u, two successive sites u^2, and k successive sites u^k. To get a cluster of size k, we need k successive occupied sites with two empty sites at the extremities, which occurs with probability

$$n_k(u) = (1-u)^2 u^k \ . \tag{1.12}$$

As described above, percolation leads to the ordered phase if we get a spanning, infinite, cluster. We thus need the probability to find a cluster with $k \to \infty$ to be non-zero. However, for any $u < 1$, the term u^k in Eq. (1.12) falls to zero for a sufficiently large k. Contrariwise, it is trivial to see how we get a single infinite cluster spanning all sites if $u = 1$. Hence, the percolation threshold and phase transition occurs at $u_c = 1$.

The behaviour of the system as it approaches this transition is particularly interesting. We rewrite Eq. (1.12) as

$$n_k(u) = (1-u)^2 u^k = (1-u)^2 \exp\left[\ln\left(u^k\right)\right] = (1-u)^2 \exp\left[k/k_\xi\right] \tag{1.13}$$

where k_ξ is the *characteristic cluster size*: $k_\xi = -1/\ln(u)$. When approaching the critical point $u_c = 1$, we use the Taylor series of the logarithm to find

$$\lim_{u\to u_c} k_\xi = \lim_{u\to 1} \frac{-1}{\ln(u)} = \lim_{u\to 1} \frac{-1}{\ln\left[u_c - (u_c - u)\right]} = \frac{1}{u_c - u} = (u_c - u)^{-1} \ . \tag{1.14}$$

[3] This formulation based on site is called *site percolation*. Another variant, called *bond percolation*, occupies the links between different sites.

While the exponent -1 is particular to the one dimensional problem, such a divergent behaviour is universal to phase transition at critical points. The characteristic size diverges at the critical point (the system becomes scale-free!) as $k_\xi = |u_c - u|^{-1/\sigma}$ where σ is called a critical exponent.

In one dimension, the spanning cluster is the sole cluster found in the ordered state. In two dimensions, the percolation threshold is found at occupation probability lower than one, such that the spanning cluster coexists with infinitely many finite clusters. This mix of small events and a single infinite event is akin to the behaviour of power-law distributed earthquake magnitudes. This raises an interesting question. If this behaviour is found in percolation only around a very precise critical value, what does this imply for the physics of earthquakes?

The paradox between the fine-tuned criticality found at a phase transition and the ubiquity of scale independence in nature is the source of ongoing debates. The next section is in fact devoted to this problem. We first present the prevalence of scale-free organizations in empirical data as well as statistical methods for their analysis. We then present an interesting hypothesis trying to explain the apparent universality of this organization across systems of diverse nature. Chapter 4 will return on the question of earthquakes and percolation when discussing the so-called self-organized criticality of systems that converge to criticality regardless of parameters and initial conditions.

1.2 Scale independence in complex systems

We now present the first series of data sets that will be used throughout the thesis: distribution of artistic and scientific productivity, distributions of sexual relations, of online activities and of word frequencies as well as the structures of the Internet and citation networks. Following the rigorous work of Clauset, Shalizi and Newman [28], we present a basic method to estimate scale exponents of these scale-independent systems.

1.2.1 Power-law distributions in empirical data

Power-law distributions in empirical data are often referred to as *Zipf's law* in honour of the linguist who gathered an impressive amount of data from social and man-made systems that all roughly followed power-law distributions [104]. While not one of the first to study these empirical distributions, he certainly brought a lot of attention to this particular class of systems and even attempted to explain their apparent universality. His explanation essentially hinges on an equilibrium condition between the effort necessary in organizing a system and and the opposite effort of analysing that organization. For instance, word frequencies in prose follow a power-law distribution because it is an equilibrium from the least effort of the writer

— use a single word for everything, a delta distribution at one end of the frequency spectrum — and the least effort of the reader — every word has a precise and unique meaning such that most words can be used only once. This is the so-called *principle of least effort* [104].

A discussion on the possible origins of these distributions will be covered in the next section. Here we will first reproduce the analysis first done by Zipf in a much more rigorous mathematical framework and using some of our own data sets. As mentioned earlier, the scale exponent γ of a power-law distribution corresponds to its slope on a log-log scale. Simple visual inspection or least-squares linear regression on the logarithm of the distribution are often used to fit the scale exponent. These methods are mostly inconsistent, due to the large scale fluctuations that occur in the tail, but are still useful to get rough estimates (e.g. whether γ is greater or lesser than two).

In this thesis, our method of choice to fit parameters is the *method of maximum-likelihood estimation* (MLE). The likelihood of a given set of parameters is given by the likelihood of the observed data given a chosen model or distribution and assuming that the set of parameters is correct. The "most likely" set of parameters is then the one for which the observed data is the most likely. The estimation problem then becomes a problem of maximizing likelihood, or equivalently its logarithm (usually simpler) noted $\mathcal{L}$.

For instance, with a sequence of observations $X = \{X_i\}$, the likelihood that X came from the normalized discrete probability distribution $P(x)$ would be the product of the likelihood of every observation: $\prod_i P(X_i)$. The log-likelihood would then simply be

$$\mathcal{L}\left(P(x)\right) = \ln\left[\prod_i P(X_i)\right] = \sum_i \ln\left[P(X_i)\right] \ . \tag{1.15}$$

In our case, since the power-law behaviour might only be present in the tail, we define the observed data as the subset of all n observations $k_i \geq k_{\min}$ where $k_{\min}$ is an appropriate bound for the power-law tail. Thus, the log-likelihood of data k_i depends on both γ and $k_{\min}$ following

$$\mathcal{L}(\gamma, k_{\min}) = \ln\left[\prod_{i=1}^{n} \frac{k_i^{-\gamma}}{\sum_{j>k_{\min}} j^{-\gamma}}\right] \tag{1.16}$$

where the sum is simply the normalizing factor of the tail. This sum can be referred to as a generalized zeta function $\zeta(\gamma, k_{\min})$. The MLE of γ and $k_{\min}$, usually noted $\hat{\gamma}$ and $\hat{k}_{\min}$, are thus obtained by maximizing

$$\mathcal{L}(\gamma, k_{\min}) = -n \ln\left[\zeta(\gamma, k_{\min})\right] - \gamma \sum_{i=1}^{n} \ln k_i \ . \tag{1.17}$$

Maximizing for γ using a given $k_{\min}$ is straightforward: we scan potential γ values, typically in steps of 10^{-2} or 10^{-3}. Choosing the right $k_{\min}$ is another story. If too low, we risk including the non-power-law behaviour (if any) of the data. If too high, we lose too many observations and the estimate of γ will also suffer. To select the correct $k_{\min}$, we simply run a range of potential values, usually from 1 to a maximal value of 10 or 100 based on some visual inspection (to avoid useless computation). For each value, we maximize the likelihood for γ. We now have somewhere between 10 or 100 candidates for $(\hat{\gamma}, \hat{k}_{\text{min}})$. We choose the one that makes the fitted power-law tail as *similar* to the original observed distribution as possible. On a finite size sample[4], we follow Clauset *et al.* and use the Kolmogorov-Smirnov distance (or KS statistic) defined as the maximum distance between the cumulative distribution function of the data $(S(k))$ and of the fitted model $(F(k))$:

$$D_{\text{KS}} = \max_{k_i \geq k_{\text{min}}} |S(k) - F(k)| \ . \tag{1.18}$$

The pair $(\hat{\gamma}, \hat{k}_{\text{min}})$ minimizing D_{KS}, among all MLE candidates, is then finally chosen.

Figure 1.2 presents results of this method on data sets which will be studied in the next sections and chapters. We note the population in each data sets as N, the distributed resource as K and the number of individuals with share $k_i = k$ as N_k. These results were organized roughly by *goodness-of-fit*. Even visually it is evident that the fits obtained for the sexual and citation networks are good only for the noisy end of the tail, whereas the fits obtained on the first three data sets are good for a significant portion of the distribution, and fail in the noise. Looking at the arXiv data, we see that the two choices could have been deemed acceptable: fit either the beginning of the distribution with a power-law or fit the noisy end.

The fact that noise usually occupies a significant portion of our distributions, again due to the large scale fluctuations that occur in the tail, is a typical problem in fitting power laws. Several methods could be considered to quantify the goodness-of-fit or the likelihood of the power-law hypothesis, typically in terms of a statistical distance similar to the KS statistic discussed above. However, we will usually refrain from doing so. First, because we will not be doing a lot of actual fits or inference of scale parameters. Second, because fitting a power-law distribution to a growing system imply that we assume that the system has reached a large enough size to be close to its asymptotic steady-state. However, we currently have no idea on how we can describe the growth of these systems or the convergence of their distribution to an actual power law. Maybe we should actually fit power-law distributions with an exponential cut-off, or any other way to account for finite size effects. With that in mind, even the distribution for the citation network in High Energy Physics (HEP) could be considered “on its way” to scale independence, but currently far from equilibrium.

For the moment, these data sets simply provide a good illustration that heavy-tailed distributions appear ubiquitously across a myriad of different human and man-made systems. Before finally delving into the modelling of these systems, we present an interesting theory to explain the origins of this organization using a statistical physics argument. Moreover, this theory will provide some insights to guide our modelling efforts.

[4] By contrast, in Chap. 3 we will need to compare fitted distributions with “infinite” solutions of analytical models. Another statistical distance will then be necessary.

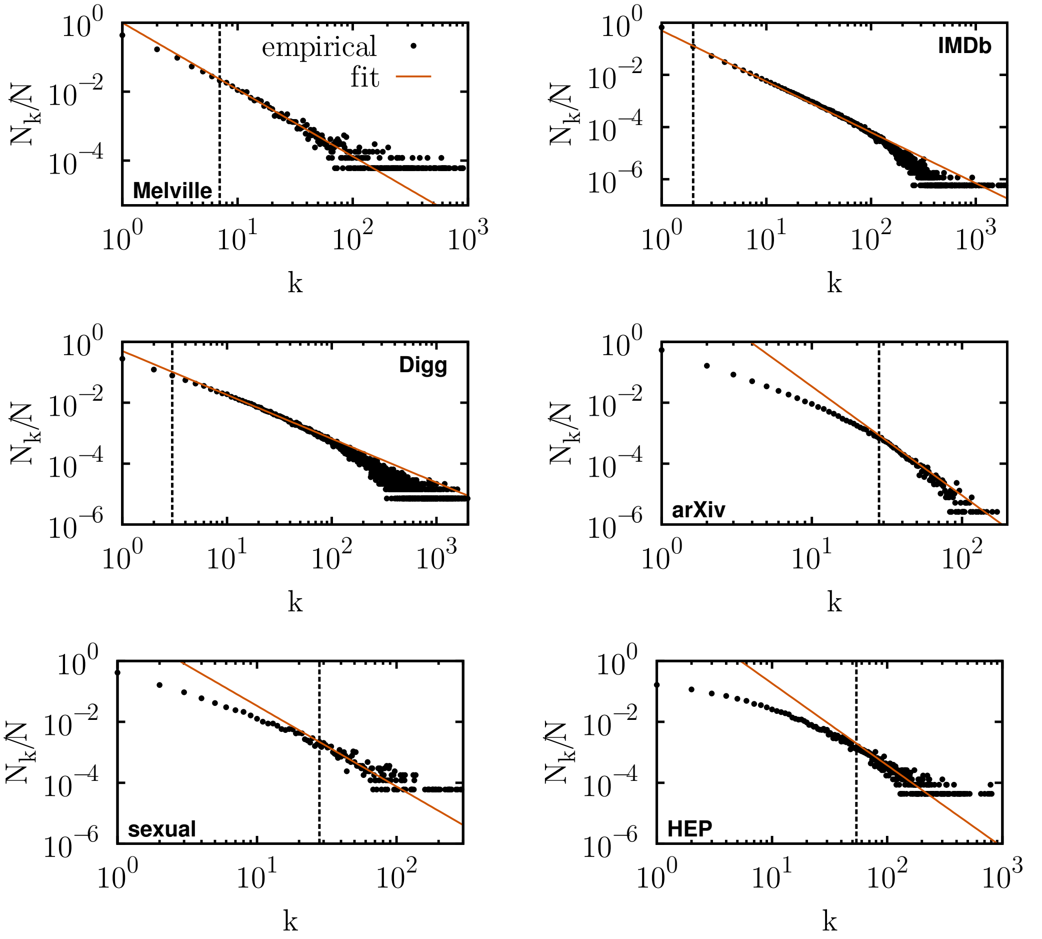


Figure 1.2 – **Fitting power-law distributions to empirical data.** Six empirical systems which may or may not follow a scale-independent distributions are fitted to a power law with the method presented in the text. (top left) Distribution of occurrences per unique word in Herman Melville's Moby Dick: unique words $N = 16,695$, total words $K = 614,680$, $\hat{k}_{\min} = 7$, $\hat{\gamma} = 1.94$. (top right) Distribution of roles per actor in the Internet Movie Database: actors $N = 1,707,525$, total roles $K = 6,288,201$, $\hat{k}_{\min} = 2$, $\hat{\gamma} = 1.95$. (middle left) Distribution of clicks per user on the Digg news aggregator website: users $N = 139,409$, total clicks $K = 3,018,197$, $\hat{k}_{\min} = 3$, $\hat{\gamma} = 1.44$. (middle right) Distribution of papers per author on the arXiv preprint archive: authors $N = 386,267$, total papers $K = 1,206,570$, $\hat{k}_{\min} = 28$, $\hat{\gamma} = 3.58$. (bottom left) Distribution of sexual relations per individual in a Brazilian escort community: individuals $N = 16,730$, total activities $K = 101,264$, $\hat{k}_{\min} = 28$, $\hat{\gamma} = 2.65$. (bottom right) Distribution of citations per paper in a subset of the High Energy Physics preprint archive: papers $N = 23,180$, total citations $K = 355,221$, $\hat{k}_{\min} = 54$, $\hat{\gamma} = 2.69$.

1.2.2 Origins of scale independence in self-organized systems

The systems presented in Fig. 1.2 have not much in common, besides their heavy-tailed distributions of course. Some concern words, others papers, others humans involved in sexual, online or professional activities. However, we aim for a universal description of their similar distributions, without any system-specific assumptions. Hence, the only common feature that we can consider is related to the fundamental statistical definition that we considered earlier for the populations under study: different elements are distinguishable only by their share of the resource. Going back to the word occurrences example: we do not distinguish words by their intrinsic definition and their usage is context free, we only differentiate them by their share of the total word count. The universal way of looking at these discrete power-law distributions is thus as a distribution of a resource (balls) within a population (bins). We now present a theory based on this simple observation: the Random Group Formation (RGF) of Baek, Bernhardsson and Minnhagen [8], an argument explaining the origins of an ubiquitous power-law distribution of balls per bin.

Random Group Formation

We have K balls to be distributed in N bins. The distribution $\{N_k\}$ of balls per bin is as yet unknown, but we know there are K "slots" for the balls to fill and the probability of throwing a ball in any slot is *a priori* equal to $P_{\text{rand}} = 1/K$. This is the Bayesian assumption of RGF. Moreover, we also know that the most likely $\{N_k\}$ will be a maximization of entropy $H\left(\{N_k\}\right) = -\sum \left(N_k/N\right) \ln \left(N_k/N\right)$. This is the statistical physics assumption of RGF.

To solve for the probability distribution N_k/N, we need to consider two constraints: $\sum N_k = N$ (population of bins) and $\sum kN_k = K$ (total resource), as well as two optimizations corresponding to our two assumptions. Our condition on entropy is already written as a maximization problem, but how can our Bayesian assumption be taken into account? We can quickly illustrate how the problem is in essence equivalent to a minimization of information cost.

All K slots are equivalent, but we are able to distinguish bins of different size. Thus, we consider the information needed to track a ball knowing the size k of the bin in which it is contained. Bins of equal size are undistinguishable, so the ball can occupy any of the kN_k slots belonging to bins of size k. The information cost of tracking one ball is then given by $\ln\left(kN_k\right)$ (in nats: a logarithmic unit of information in natural base). The information cost corresponding to the distribution $\{N_k\}$ is thus $I\left(\{N_k\}\right) = \sum N_k \ln\left(kN_k\right)/N$, which we seek to minimize. This approach by information cost minimization is due to the theoretical lexicography[5] background of the authors. We can also offer a physical point of view on their

[5] Theoretical lexicography is the study of relationships within the vocabulary of a language.

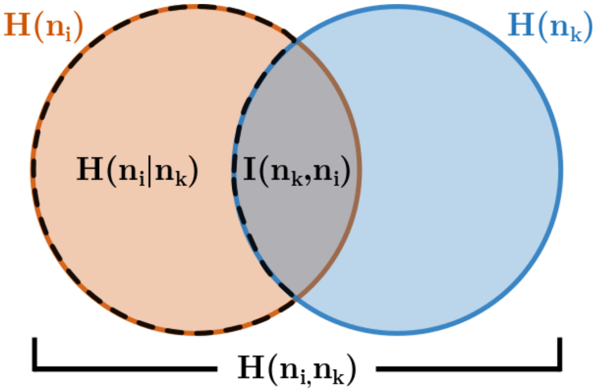


Figure 1.3 – **Diagram of mutual information based on joint and constrained entropies.** If $\{n_i\}$ is the probability of finding a given ball in a given slot and $\{n_k\}$ is the slots per bin distribution, then the mutual information $I\left(n_k, n_i\right)$ is the overlap of their respective marginal entropies. Consequently, it is the reduction in the conditional or constrained entropy $H\left(n_i|n_k\right)$ caused by our knowledge of the bin size distribution $\{n_k\}$. This constrained entropy is illustrated as a crescent moon. This figure was adapted from the Wikimedia Commons file "Image:Entropy-mutual-information-relative-entropy-relation-diagram.svg".

argument based on information theory [29].

Consider the distribution of balls per bin, $\{n_k\}$, and the probability distribution of finding a given ball in a given slot i, $\{n_i\}$. Let us assume that the entropy of the distribution of slots per bin $\{n_k\}$ is fixed by the first optimization. We then seek to maximize the constrained entropy $H\left(n_i|n_k\right)$ for the probability of finding a given ball in a given slot ($\{n_i\}$) given $\{n_k\}$ and considering the distinguishability of bins of different size. Hence, the entire theory rests on a problem of *constrained entropy maximization*: we maximize both the entropy of the distribution of bin size $\{n_k\}$ and of ball localization $\{n_i\}$ given $\{n_k\}$.

The conditional entropy is also expressed as $H\left(n_i|n_k\right) = -I\left(n_k, n_i\right) + H\left(n_i\right)$ where $I\left(n_k, n_i\right)$ is a mutual information term and where the entropy $H\left(n_i\right)$ of $\{n_i\}$ regardless of $\{n_k\}$ is known through our Bayesian assumption that slots are equiprobable. Thus, maximizing $H\left(n_i|n_k\right)$ is equivalent to minimizing the mutual information term $I\left(n_k, n_i\right)$. More conceptually, the information term to be minimized is equivalent to the reduction in uncertainty on the occupied slot i due to our knowledge of the size k of its bin. These equivalences are illustrated in Fig. 1.3.

We can show that the mutual information is equivalent to the information cost discussed in the previous analysis. We use the distributions $P(n_k) = N_k/N$ (by definition), a uniform $P(n_i)$ (by assumption) and $P(n_k, n_i) = (N_k/N)(kN_k)^{-1} = (kN)^{-1}$ which means that we first select a bin size (first term) and then a slot within the bins of that size (second term). The definition of mutual information (see [29] §2.4) is

$$I\left(n_k, n_i\right) = \sum_k \sum_{i|k} P(n_k, n_i) \ln\left[\frac{P(n_k, n_i)}{P(n_k)P(n_i)}\right] \tag{1.19}$$

and it can be simplified by using the definition of joint probability distribution (often written

$P_{x,y}(x,y) = P_{x|y}(x|y)P_y(y)$ in probability theory)

$$I\left(n_k, n_i\right) = \sum_k \sum_{i|k} P(n_k, n_i) \ln \left[\frac{P(n_i|n_k)}{P(n_i)}\right] \ . \tag{1.20}$$

Given a size k, $P(n_i|n_k)$ is simply kN_k/K, such that $P(n_i|n_k)/P(n_i) = kN_k$. The last expression thus reduces to

$$I\left(n_k, n_i\right) = \sum_k \sum_{i|k} \frac{1}{kN} \ln\left(kN_k\right) = \sum_k \frac{N_k}{N} \ln\left(kN_k\right) = I\left(\{N_k\}\right) \ . \tag{1.21}$$

We now have a double optimization problem, for entropy and mutual information (or information cost), constrained by the number of balls and bins. Instead of straightforwardly solving the problem, Baek *et al.* use some mathematical sleight of hand. Once again, we assume that we know the entropy $H(\{N_k\})$, when in fact we will be able to fix it *a posteriori* by adding a third constraints from the system. This allows us to remove one of the function to be optimized. The remaining optimization problem can then be written as the following Lagrange function

$$F\left(\{N_k\}\right) = I\left(\{N_k\}\right) + \lambda_1 \left| N - \sum_k N_k \right| + \lambda_2 \left| K - \sum_k kN_k \right| + \lambda_3 \left| H + \sum_k \frac{N_k}{N} \ln \frac{N_k}{N} \right| \ . \tag{1.22}$$

The solution to Eq. (1.22), obtained by solving $\partial F\left(\{N_k\}\right)/\partial N_k = 0$ for N_k, is given by

$$\frac{N_k}{N} = A \exp\left(-bk\right) k^{-\gamma} \tag{1.23}$$

with $A = \exp\left[-1 - \lambda_1/(1-\lambda_3)\right]$, $b = \lambda_2/(1-\lambda_3)$ and $\gamma = (1-\lambda_3)^{-1}$ where the multipliers λ_i must be fixed with the three constraints. On the one hand, we have accomplished our goal. The very general assumptions made on the balls in bins system do lead to power-law behaviour for the distribution of balls per bin. The exponential cut-off, i.e., the factor $\exp\left(-bk\right)$, is simply related to the finite size of the system, i.e., $\sum kN_k = K$. On the other hand, we still have a floating constraint relating to the entropy H of the system.

This problem can be solved using the logic we first applied to the size of the largest earthquake in Sec. 1.1.2. First consider the following dependency of H on the scale exponent γ,

$$\begin{aligned} H &= A \sum \exp\left(-bk\right) k^{-\gamma} \left[bk + \gamma \ln k\right] - \ln A \\ \frac{\partial H}{\partial \gamma} &= -A \sum \exp\left(-bk\right) k^{-\gamma} \ln k \left[bk + \gamma \ln k - 1\right] < 0 \ , \end{aligned} \tag{1.24}$$

meaning that a larger scale exponent (making the distribution less broad) implies a loss of entropy. In other words, a smaller entropy corresponds to a smaller value of $k_{\max}$. This is a well defined value for any data set, but how does it close our theoretical solution? We simply swap the constraint on entropy by a constraint on the related value $k_{\max}$. To this end, remember that the largest expected value when sampling a power-law tail is defined by the point k_c above which only a single data point is expected. Using once again the cumulative distribution function $C(k)$, this point is given by $C(k_c) = 1/N$, with an expected size $\langle k_{\text{max}} \rangle = \sum_{k=k_c}^{\infty} kN_k/N$ for the actual k_{max} obtained from the data. Our constraints can then be fixed from any data set through: K (number of balls), N (number of bins), and k_{max} (largest number of balls in a single bin).

More precisely, to obtain a final solution, the parameters of Eq. (1.23) are solved self-consistently with the three following conditions:

$$\sum_{k=1}^{K} A\frac{e^{-bk}}{k^{\gamma}} = 1 \tag{1.25}$$

$$\sum_{k=1}^{K} Ak\frac{e^{-bk}}{k^{\gamma}} = K/N \tag{1.26}$$

$$\left[\sum_{k=k_c}^{K} Ak\frac{e^{-bk}}{k^{\gamma}}\right]\left[\sum_{k=k_c}^{K} A\frac{e^{-bk}}{k^{\gamma}}\right]^{-1} = \langle k_{\text{max}} \rangle \; . \tag{1.27}$$

This set of equations is solved straightforwardly. We assume a scale exponent γ and solve for b with the ratio of the first two equations, then normalize the distribution to fix A. We repeat this procedure until we converge to the value of γ leading to a distribution respecting the last equation.[6]

Figure 1.4 revisits the data sets studied in Fig. 1.2, but now fitted by solving the RGF equations. While it is up for discussion whether these distributions are more faithful to the empirical data than the purely power-law fit, it should be clear that RGF at least captures some of the finite size effects present in the data sets. Unfortunately, RGF still relies on the assumption that we are at least close to equilibrium as it is based on an entropy optimization framework. A comparison between the statistical approach of Clauset *et al.* and the theoretical approach of Baek *et al.* is given in Table 1.1. The methods differ mostly when the statistical approach fits the power-law behaviour to what RGF considers as part of the finite-size cut-off, as evident in the arXiv, sexual and HEP data.

Finally, even if the Random Group Formation theory might tell us *why* systems tend to organize in a scale-independent manner, it does not tell us *how* they reach that organization. Obviously, these data sets are all of finite size and the systems they describe are growing and evolving such that both the average K/N and $\langle k_{\text{max}} \rangle$ are changing in time. Can we look at the distribution at a given point in time and predict what it might look like in the future? How do individual bins tend to grow? Can we guess how likely it is for a rare word to reappear if a lost second epilogue to Moby Dick is discovered? To answer these questions, we must already turn away from optimization problems as they imply a certain equilibrium, and turn to a description in terms of stochastic growth processes.

[6] A Newton-Raphson method can be applied to make the solution converge while maintaining consistency with Eqs. (1.25) and (1.26).

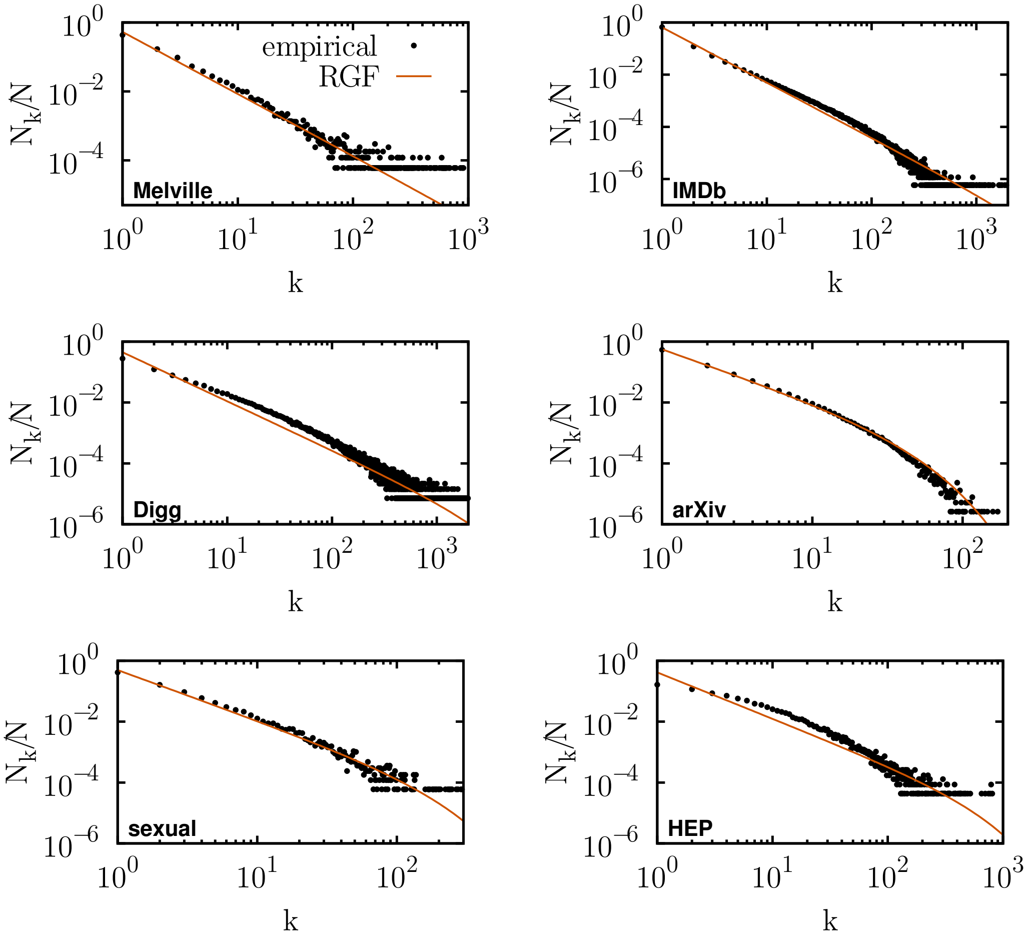


Figure 1.4 – **Applying Random Group Formation to empirical data.** Data sets used in Fig. 1.2 are now analysed with the Random Group Formation theory. The fits now take finite size into account and are parametrized as follows. (top left) Melville's Moby Dick: $k_{\max} = 14,176$, scale exponent $\gamma = 1.80$ and exponential cut-off $bK = 21.95$. (top right) Internet Movie Database: $k_{\max} = 6,059$, scale exponent $\gamma = 2.11$ and exponential cut-off $bK = 1,329.10$. (middle left) Digg: $k_{\max} = 10,526$, scale exponent $\gamma = 1.61$ and exponential cut-off $bK = 1052.28$. (middle right) arXiv: $k_{\max} = 214$, scale exponent $\gamma = 1.72$ and exponential cut-off $bK = 38,938.76$. (bottom left) sexual data: $k_{\max} = 615$, scale exponent $\gamma = 1.66$ and exponential cut-off $bK = 665.89$. (bottom right) High Energy Physics: $k_{\max} = 2,414$, scale exponent $\gamma = 1.52$ and exponential cut-off $bK = 643.02$.

Table 1.1 – Summary of database sizes, quantities and estimated parameters.

System	Population N	Resource K	$\hat{\gamma}$	γ_{RGF}
Melville	$16,695$ unique words	$614,680$ written words	1.94	1.80
IMDb	$1,707,525$ actors	$6,288,201$ roles	1.95	2.11
Digg	$139,409$ users	$3,018,197$ clicks	1.44	1.61
arXiv	$386,267$ authors	$6,288,201$ roles	3.58	1.72
sexual	$16,730$ individuals	$101,264$ activities	2.65	1.66
HEP	$23,180$ papers	$355,221$ citations	2.69	1.52

1.2.3 Link with growth processes

We now wish to start modelling the growth of scale-independent systems. From this point onward, we will only put balls in bins as in the analogy considered in the previous section. We first simply seek a model whose assumptions agree with RGF and whose steady state is similar to Eq. (1.23).

In one of their previous publications [79], the authors behind the RGF theory proposed an answer. While Minnhagen *et al.* originally tackled the problem of population distribution in different towns, we reword their contribution as a scheme to distribute K balls amongst N bins. We initially start with any distribution of balls per bin. We then simply pick two balls at every time step and move one of them to the other's bin. Therefore, a bin of size k can either win or lose a ball at every time step, both with probability $\frac{1}{2}k/K$.

Using methods similar to those we will introduce in the next chapter, we can show that the steady-state solution of this simple scheme is indeed given by Eq. (1.23). Moreover, the model not only produces a power-law distribution with exponential cut-off, but also follows the same hypotheses as RGF. Firstly, bins are only distinguishable by their size, meaning that two bins of the same size have the same probabilities of winning or losing a ball. Secondly, the process maximizes entropy and is ergodic [79] (visiting all possible states), such that it respects both the statistical physics and Bayesian assumptions of RGF.

To invent such a simple process to generate power-law distributions, Minnhagen *et al.* took inspiration from one of the first model of scale-independent systems which was proposed by Gibrat. While originally a description of the growth of an economical firm [47], we can also reword Gibrat's model in terms of balls and bins. Simply put, we follow the number of balls k in a single bin and add or remove balls proportionally to their current number k. We should highlight at this point that this rule of proportionality is the key ingredient shared by both models and will be an important mechanism for the rest of this thesis.

However, there are several problems with Gibrat's model which is why we will refrain from focusing too much on it. Firstly, we are modelling an independent bin, whereas in reality if a ball goes into a given bin, it is not going in any other. This means that that the probability of a given word to be the next one used in a text is not only dependent on its own intrinsic usage, but also on the usage of all other possible words. Secondly, and this applies also to the model of Minnagen *et al.*, the empirical data considered thus far concern quantities that are never decreasing — total occurrences in a text, total sexual activities, etc. — whereas we here allow balls to be removed from the bin. Thirdly, and more importantly, both model concerns static systems where both the number of balls K and the number of bins N are fixed in time. We want to consider growth processes that can describe systems as they evolve towards their asymptotic state.

In short, our model must consider interdependent elements, and our model must be a growth process and not only a scheme of distributing balls in bins. Models of the sort make perfect sense for the systems hitherto considered. All quantities are discrete just like counting balls, monotonously increasing as we never erase words or remove balls and growing as words are created and new individuals join the systems. The next chapter covers a family of processes which respects these criteria.

Chapter 2

On preferential attachment

Résumé

Nous lançons maintenant des balles dans des urnes. Ce chapitre couvre une famille de processus stochastiques de croissance ayant beaucoup en commun avec le modèle de Gibrat présenté à la fin du précédent chapitre. En fait, ils utilisent tous la même idée centrale: l'attachement préférentiel. Ces processus seront étudiés en détails dans le reste de la thèse. Nous discutons ici de leur invention, de manière indépendante, par Yule [102] et Simon [92]. Pour le premier, il s'agit d'un modèle d'évolution pour expliquer la distribution d'espèces observées par genre biologique. Plus un genre contient d'espèces, plus il risque d'y avoir des mutations menant à de nouvelles espèces du même genre. Pour le second, il s'agit d'un modèle jouet où des balles sont lancées dans des urnes préférentiellement au nombre de balles qu'elles contiennent. Tous deux utilisent donc la même logique sous-jacente de proportionnalité: les riches s'enrichissent.

Nous présenterons également certaines re-découvertes de l'attachement préférentiel dont celle de Barabási et Albert dans le cadre de la science des réseaux [12]. Au fil du chapitre, nous introduisons et étudions l'approche par équation maîtresse qui sera au cœur de toutes nos analyses. Nous terminons avec des études détaillées de l'attachement préférentiel — sur une propriété intéressante de notre solution explicite [57], sur l'effet de la loi de proportionalité [67, 65] et sur l'influence du taux d'introduction d'urnes [103] — qui permettent d'illustrer la flexibilité des équations maîtresses. Les résultats obtenus dans ces études pavent la voie vers une de nos contributions principales présentée au chapitre suivant.

Summary

We now throw balls in bins. This chapter covers a family of stochastic growth processes which share some properties of the Gibrat model discussed at the end of the previous chapter. In fact, they all share the same basic principle: preferential attachment. These processes will be studied in great details for the remainder of the thesis. Here, we first cover their independent invention by Yule [102] and Simon [92]. For the former, the process is a model of evolution explaining the observed distribution of species in biological genera. The more species there are in a genus, the more likely it is that mutations will lead to a new species of the same genus. For the latter, the process is a toy model where balls are thrown in urns preferentially to the number of balls they already contain. Both draw upon the same underlying logic of proportionality: the rich get richer.

We also present certain rediscovery of preferential attachment including the Barabási and Albert variant of the process introduced in network science [12]. In presenting these models, we introduce and study the master (or rate [1]) equation approach which lies at the heart of all our analyses. We then end with detailed study of preferential attachment — on an interesting property of our explicit solution [57], on the effect of the proportionality rule [67, 65] and on the influence of the urn introduction rate [103] — which allow us to illustrate the flexibility of the master equations. The results obtained in these studies pave the way towards one of our main contributions presented in the following chapter.

2.1 The history of preferential attachment

2.1.1 Yule's process

One of the first [2] classic cases of observation of scale independence in empirical data is due to Willis and Yule in 1922 when they analysed the size distribution (in number of species) of biological genera [100, 99]. Their study, somewhat surprisingly, highlighted the particular shape of that distribution: a power law (see Fig. 2.1 for one of the first published figures presenting power-law distributed empirical data as a straight line on logarithmic scales).

In a subsequent publication [102], Yule proposed a mathematical model of evolution to explain the observed distribution. In biological terms, the model is defined as follows. Two types of mutations are possible: *specific mutations*, which produce a new species of the same genus and *generic mutations*, which produce a new species in a new genus. These mutations occur respectively within each species at a rate s and within each genus at a rate g. Yule then

[1] In this thesis, we can mostly use both *rate equation* and *master equation* interchangeably.
[2] After a first observation by Estroup in 1916 [44].

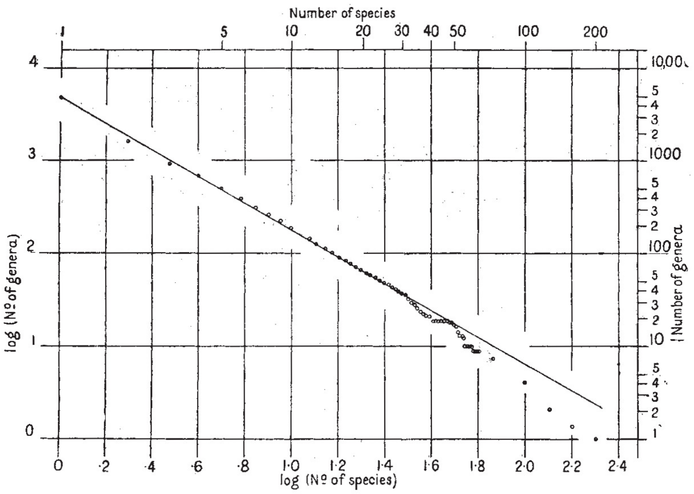


Figure 2.1 – **Number of species per genus of flowering plants circa 1922.** The straight line represents a perfect power law as the data are presented under logarithmic scales. Reproduced with permission [99].

described the mean total number of genera at time t, $N(t) = \exp(gt)$, and the age distribution of these genera,

$$n_a = g \exp(-ga) \ , \tag{2.1}$$

as well as the average size (number of species) for a genus of age a,

$$k(a) = \exp(sa) \ . \tag{2.2}$$

to show how the size distribution n_k asymptotically falls as $n_k(t \to \infty) \propto k^{-(1+g/s)}$. To simplify his analysis, we can invert Eq. (2.2) to obtain

$$a(k) = \frac{1}{s} \ln k \ ; \quad da = \frac{1}{s}\frac{dk}{k} \ . \tag{2.3}$$

From this relation, we can translate the age distribution (i.e., Eq. (2.1)) to a size distribution following $n_k = |J| g \exp(-ga(k))$, where $|J|$ is the Jacobian from the substitution of age to average size given by Eq. (2.2). We thus find

$$n_k = \frac{g}{s} k^{-(1+\frac{g}{s})} \tag{2.4}$$

which does indeed follow $n_k \propto k^{-\gamma}$ with a scale exponent $\gamma = 1 + g/s$. Finally, analysing the data presented in Fig. 2.1 yields $\gamma \simeq 1.5$ such that $s \simeq 2g$.

2.1.2 Simon's model

It is two decades later that Zipf published his results on the power-law distributions observed in various spheres of human activities: distributions of word occurrences in prose, of train cargoes between cities, of city populations, of time interval between consecutive repetitions of notes in Mozart's Bassoon Concerto in B-flat Major, of sales of diverse goods among others [104]. The ubiquity of this distribution captured the interest of Simon, then mostly an economist and political scientist, who was not entirely convinced by Zipf's interpretation (his principle of least effort).

Simon then developed a stochastic process similar in principle to Yule's model of evolution, although much more general in nature, and clearer in its assumptions and mechanisms [92]. The hope being to explain the omnipresence of scale independence regardless of the systems' intrinsic nature and mechanisms. In fact, Simon does not postulate or assume anything about the system under study, with the sole exception of one mechanism: *the rich get richer*. Mathematically, we can state that the chances (or the rate) of growth for an element of the system are directly proportional to its "size." Here, size can refer to an actual size (like the number of species in a genus in Yule's process) or to a share of a given resource or to the number of past occurrences of a given event. Notice that this rule was implicitly included in Yule's process in the exponential growth assumed for each genus.

In short, Simon's process is a simple urn model in which balls are thrown in urns via a preferential scheme: urns with x balls are x/y times mores likely to get the new ball than urns with y balls. These balls can represent a new word written in a text, money being invested, or the publication of new scientific papers. The urns would then represent unique words, financial firms, or scientists. In all cases, the probability that a ball falls in an urn already containing k balls is simply given by k normalized by the total number of balls previously thrown (i.e., the content of all other urns). To allow the population (the number of urns) to grow, at each throw, the ball has a fixed probability p of falling in a new urn.

Simon analysed his model with rate equations. Hereafter, this is the sole approach we will use to follow the time evolution of any stochastic process. The idea is to follow the evolution of the average state of the system, allowing for whatever relevant heterogeneity. In this case, we want to follow the average absolute number $N_k(t)$ of urns containing k balls after t throws. We thus write

$$N_k(t+1) = N_k(t) + \left[(1-p)\frac{(k-1)N_{k-1}(t) - kN_k(t)}{t} + p\delta_{k,1}\right] . \tag{2.5}$$

Let us take a second to explain how these equations are constructed. It is essentially a finite difference equation (as time is discrete). Hence, $N_k(t+1)$ is equal to its previous state plus the average variation due to the event during that time step. In this case, the event can be a birth or a growth. Birth events occur with probability p and affect only the urns with $k = 1$ (hence the Kronecker delta $\delta_{k,1} = 1$ if $k = 1$ and 0 otherwise) and have a weight of one (a single urn is born). Growth events occur with complementary probability $1 - p$ and affect urns with k balls negatively if one of them is chosen (probability $kN_k(t)/t$) or positively if an urn with $k - 1$ balls is chosen (probability $(k - 1)N_{k-1}(t)/t$). For consistency, we set the boundary condition $N_0(t) = 0$ for all t.

When t becomes very large, such that the time step is much smaller than t itself, we can switch to a continuous time process to simplify the upcoming analysis [3]:

$$N_k(t+dt) = N_k(t) + dt\left[(1-p)\frac{(k-1)N_{k-1}(t) - kN_k(t)}{t} + p\delta_{k,1}\right] . \tag{2.6}$$

We then transform Eq. (2.6) in an equation for the proportion $n_k(t)$ of urns containing k balls at time t. To do so, we use the fact that $\{N_k(t)\}$ is simply $\{n_k(t)\}$ multiplied by the average number $N(t)$ of urns at time t, i.e. pt. We thus write

$$p(t+dt)n_k(t+dt) = ptn_k(t) + dt\Big\{p\,(1-p)\left[(k-1)\,n_{k-1}(t) - kn_k(t)\right] + p\delta_{k,1}\Big\} \tag{2.7}$$

from which we obtain an ordinary differential equation (ODE) of the form

$$\lim_{dt\to 0}\frac{(t+dt)n_k(t+dt) - tn_k(t)}{dt} = \frac{d}{dt}\left[tn_k(t)\right] = (1-p)\left[(k-1)\,n_{k-1}(t) - kn_k(t)\right] + \delta_{k,1} . \tag{2.8}$$

The steady-state ensemble $\{n_k^*\}$ (defined by $\frac{d}{dt}n_k(t) = 0\ \forall\ k$) can straightforwardly be written through

$$\frac{d}{dt}\left[tn_k(t)\right] = n_k(t) + t\frac{d}{dt}n_k(t) = (1-p)\left[(k-1)\,n_{k-1}(t) - kn_k(t)\right] + \delta_{k,1} \tag{2.9}$$

which yields

$$n_k^* = (1-p)\left[(k-1)\,n_{k-1}^* - kn_k^*\right] + \delta_{k,1} \tag{2.10}$$

or

$$n_k^* = \frac{(1-p)\,(k-1)\,n_{k-1}^* + \delta_{k,1}}{1 + k\,(1-p)} . \tag{2.11}$$

By induction, we get

$$n_k^* = \frac{\prod_{m=1}^{k-1} m\,(1-p)}{\prod_{m=1}^{k}\left[1 + m\,(1-p)\right]} \quad \forall\ k > 1 \tag{2.12}$$

such that

$$\frac{n_k^*}{n_{k-1}^*} = \frac{(k-1)\,(1-p)}{1 + k\,(1-p)} . \tag{2.13}$$

We can then show that this last relation decreases as a power law for $k \gg 1$,

$$\lim_{k\to\infty}\frac{n_k^*}{n_{k-1}^*} = \left(\frac{k}{k-1}\right)^{-\gamma} , \tag{2.14}$$

with

$$\gamma = \lim_{k\to\infty}\left\{\log\left(\frac{(k-1)(1-p)}{1+k(1-p)}\right) \Big/ \log\left(\frac{k-1}{k}\right)\right\} = \frac{2-p}{1-p} . \tag{2.15}$$

Simon's model thus produces systems featuring a scale-independent distribution of some quantity (balls) per element (urns) with a scaling exponent $\in [2,\infty)$. It is a simple matter to simulate Simon's model (Monte Carlo experiment) for instance to reproduce the distribution of word occurrences in a text, see Fig. 2.2. This both confirms the efficiency of the model itself, and of our rate equation analysis [4].

[3] The effect of the continuous time approximation will be studied in Sec. 2.2.1.

[4] This analysis could also be called a mean-field analysis, as the rate equation determines the time evolution of the average event only.

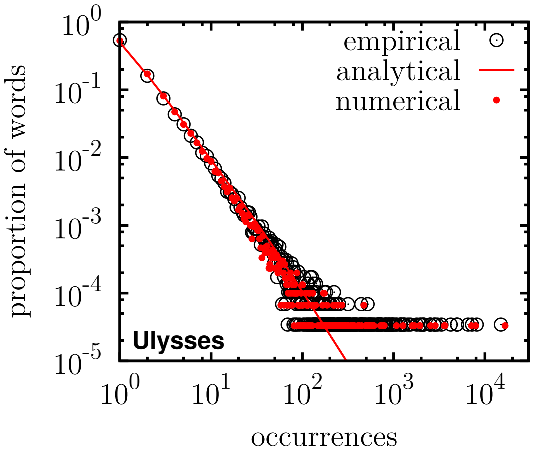


Figure 2.2 – **Word occurrences in James Joyce's Ulysses with Simon's model.** The Monte Carlo results are obtained from a single realization of Simon's model, with the parameter p fixed empirically: $p \simeq 0.112$, equal to the ratio of the number of unique words (30 030) to the total word count (267 350). The analytic curve is obtained by integrating Eq. (2.6).

2.1.3 Solution of Simon's model with Yule's method

While the rate equation approach is useful mainly to get a temporal description, the analysis used to describe Yule's process is almost undeniably simpler. It is therefore interesting to reproduce this analysis with Simon's model and, in doing so, highlight the link between the two.

If we are interested, for example, in a given urn containing k of the t balls, we can write the probability that the next ball falls into the same urn as

$$P\left[k \rightarrow k+1\right] = (1-p)\frac{k}{t} \tag{2.16}$$

which is equivalent to

$$P\left[k \rightarrow k+1\right]dt = (1-p)\frac{k}{t}dt \tag{2.17}$$

with $dt = 1$. By introducing an arbitrary time $\tilde{t} = \ln t$ (which is a continuous variable in the limit $t \rightarrow \infty$) we can write

$$P\left[k \rightarrow k+1\right]dt = (1-p)kd\tilde{t}\ . \tag{2.18}$$

Similarly, we can write the probability that the next ball falls in a new urn, that is the probability of a birth event in the population: $N \rightarrow N+1$,

$$P\left[N \rightarrow N+1\right]dt = pdt = ptd\tilde{t}\ . \tag{2.19}$$

We are now essentially back at the starting point of analysis of Yule's process (with urns ↔ genera and species ↔ balls). The only difference between the two processes being that the "generic mutation rate" of Simon's model is directly proportional to the number of "species", $\tilde{g} \equiv g + s$, as we lost a degree of freedom. We now have $g \leftrightarrow p$ and $s \leftrightarrow (1-p)$, such that $\tilde{g} = g + s = 1$. With the new continuous time, Eq. (2.19) implies $N(t) = \exp(\tilde{t})$ such that the age distribution of urns can be written as

$$n_a = p \exp(-\tilde{g}\tilde{t}) = p \exp(-\tilde{t}) \ , \tag{2.20}$$

and the average number of balls in an urn of age a,

$$k(a) = \exp\left[(1-p)\tilde{t}\right] \ . \tag{2.21}$$

These two expressions are equivalent to Eqs. (2.1) and (2.2). Repeating the last steps of our previous analysis, we once again find that the scale exponent produced by Simon's model must be

$$\gamma = 1 + \frac{1}{1-p} = \frac{2-p}{1-p} \ . \tag{2.22}$$

Inversely, we could use the rate equation approach to analyse more completely Yule's process [78]. It then becomes clear that the two processes are in fact two versions of the same basic model, one with discrete (Simon) and one with continuous (Yule) time. The fact that Simon's model has one less degree of freedom manifests itself in the range of reproducible scale exponent, going from 1 to infinity in Yule's process and only starting at 2 in Simon's case. The physical intuition behind this difference being that in Simon's model the average number of balls per urn must always be $t/pt = 1/p$ whereas it can diverge in Yule's process. As we saw in the first chapter, scale-independent organizations with diverging mean imply a scale exponent $\gamma \leq 2$. This distinction will become more manifest in the next chapter.

2.1.4 Network growth through preferential attachment

Probably because of its simplicity, and the ubiquity of scale independence, Yule's process has been re-invented multiple times and variations of it are now found in almost any imaginable science. For instance, Champernowne's model was developed to explain Pareto's law for the capital of financial firms and is similar in spirit to Yule's original process [24]. The first study of this type by a physicist is probably the work of de Solla Price who proposed to explain scale independence in bibliometric (essentially citations network) by a process which he coined *cumulative advantage* [32]. This process can be summarized as follows: popularity is popular. That is to say, the more an article has been cited in the past, the more likely it is to be cited in the future.

We must then wait roughly thirty years before the process was once again re-invented, in the context of what is now called *network science*, by Barabási and Albert who then coined the term *preferential attachment* (PA) [12]. Their study looked at the degree distribution, i.e., the distribution of links per node, which featured a power-law tail in a fraction of the World-Wide Web, in a portion of the American power grid and in a collaboration network between actors. They thus proposed the following stochastic model of network growth based on preferential attachment. At each time step dt, a new node is added to the system and connected to m existing nodes, randomly selected but proportionally to their current degree. We can use our previous formalism to write down rate (or master) equation to follow the number of nodes $N_k(t)$ with degree k at time t. Following the logic previously applied to Simon's model, we write

$$N_k(t+dt) = N_k(t) + dt\left\{m\frac{(k-1)N_{k-1}(t) - kN_k(t)}{2mt} + \delta_{k,m}\right\} . \tag{2.23}$$

Note that the normalization factor is now $2mt$, which is the total number of degree (half-link) at time t, since every time step introduces $2m$ resources (degree) to complete m links. Also, we neglected the probability to select the same node more than once within one time step as these terms all vanish at least as fast as $1/t$ [36]. We can simplify Eq. (2.23),

$$N_k(t+dt) = N_k(t) + dt\left\{\frac{1}{2}\frac{(k-1)N_{k-1}(t) - kN_k(t)}{t} + \delta_{k,m}\right\} , \tag{2.24}$$

to fall back on Simon's model with $p = 1/2$. From our previous analysis, we already know that the scale exponent of the asymptotic steady-state distribution should be $\gamma = (2-p)/(1-p)$, such that the degree distribution of the Barabási-Albert's model (BA) follows $N_k(t) \propto k^{-3}$.

The sole parameter of the BA model is the initial degree m of new nodes which only fixes a minimal bound for the distribution without the scale exponent of its tail. To widen the applicability of the model, we could take inspiration from Simon and rather than creating a new node at each time step, allow the model to create links between two existing nodes. We will fix $m = 1$, and introduce a new node only with probability p, while with complementary probability $1-p$ we will create the link between two old nodes. The distribution should now follow

$$N_k(t+dt) = N_k(t) + dt\left\{ (2-p)\,\frac{(k-1)N_{k-1}(t) - kN_k(t)}{2t} + p\delta_{k,1}\right\} , \tag{2.25}$$

which is the model that we use to reproduce the degree distribution of the World-Wide Web in Fig. 2.3. Our last equation slightly differs from Eq. (2.5), the term $(2-p)$ is used because with probability p we give only one half-link to an existing node (to connect to the new one) and with probability $1-p$ we give them two half-links such that the average is $p+2(1-p) = 2-p$.

2.2 Detailed study of classical preferential attachment

The mathematics will now get a little more involved as we study preferential attachment processes in greater detail. However, the techniques will remain somewhat similar since, as seen with the two previous models, the rate (or master) equation approach is perfectly suited for preferential attachment processes.

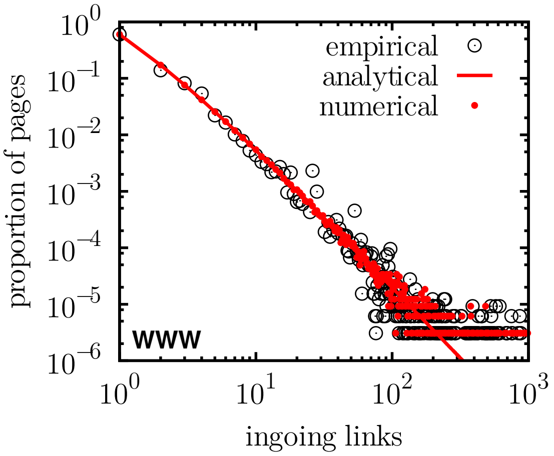


Figure 2.3 – **Degree distribution on web pages of the Notre-Dame University domain.** Reproduction of the degree distribution found in a subset of the Word-Wide Web through a generalization of the Barabási-Albert model. The numerical results are a single Monte Carlo realization of the process and the analytical results is obtained by integrating Eq. (2.25).

Let us now present the most general model that we will hereafter use. A preferential attachment process will distribute $K(t) = \sum_k kN_k(t)$ balls (or units of a resource) amongst $N(t)$ urns (or individuals). We can follow the average absolute number $N_k(t)$ of urns with k balls at time t using a finite difference equation of the form

$$N_k(t+1) = N_k(t) + [1-p(t)]\,\frac{G(k-1)N_{k-1}(t) - G(k)N_k(t)}{\sum_j G(j)N_j(t)} + p(t)\delta_{k,1} \tag{2.26}$$

where $p(t)$ is a time dependent birth function (probability $p(t)$ of a birth at the t-th time step) and $G(k)$ is a general growth function ($G(k) = k$ being the classical linear preferential attachment).

The next few sections concern, in order: the time solution of the classical preferential attachment process [57], the steady-state solution with different attachment kernels ($G(k)$) [67, 65] and the approximated solution with a time-dependent birth rate [103]. In the interest of also providing a wide overview of the methods involved, each of these three sections will also introduce different methods of solution. More importantly, the first section leads to a result hinting at the importance of time-dependent birth rate in preferential attachment; the second section demonstrates how *linear* preferential attachment is actually a critical process separating a condensate state (supra-linear attachment) and an homogeneous state (sub-linear attachment); and the third section presents how values of $\gamma < 2$ can be obtained with time-dependent birth rate. These results pave the way to one of our main contribution, which will be presented in the following chapter.

2.2.1 Time evolution of constant linear preferential attachment

The time evolution of a preferential attachment process is usually obtained simply by iterating the discrete time rate equations. However, it can be useful to have an explicit solution in continuous time, either to speed up the calculations or simply to compare the difference between the two processes. We will solve Simon's model explicitly under the approximation of continuous time. We thus assume $G(k) = k$ and $p(t) = p$. Note that the results presented are already published [57], and that our analysis uses some methods first introduced by Morin [74].

2.2.1.1 Discrete versus continuous time processes

The transition to continuous time simply implies that p now refers to a birth rate, as opposed to a birth probability within a discrete time step. The corresponding rate $1 - p$ thereby corresponds to the growth rate of existing elements. This means that in a given time interval $[t, t + dt]$, this new stochastic process could create an infinite number of elements with probability $\lim_{dt\to 0} (pdt)^{1/dt}$, whereas the discrete version could only create one element with probability p. While it is highly improbable that continuous time PA results in a system several orders of magnitude larger than pt, as the average will be exactly pt, there is no maximal size per se.

This sort of continuous time dynamics is better described using simple ODEs. To this end, we once again follow N_k, the number of bins with k balls. Using the method of Sec. 2.1.2, we directly write the following ODE

$$\frac{d}{dt}N_k(t) = \rho\,\delta_{km} + R_{k-1}(t)N_{k-1}(t) - R_k(t)N_k(t) \tag{2.27}$$

where ρ is the birth rate, m is the size of new bins (e.g. to include the Barabási-Albert model) and $R_i(t)$ is the attachment rate on entities of size i, which we define using a growth rate κ, an initial total size m_0 and a normalization rate λ:

$$R_i(t) = \frac{\kappa i}{m_0 + \lambda t} \ . \tag{2.28}$$

It proves useful to rewrite (2.27) in dimensionless form as

$$\frac{d}{d\tau}N_k(\tau) = \overline{\rho}\,\delta_{km} + \overline{R}_{k-1}(\tau)N_{k-1}(\tau) - \overline{R}_k(\tau)N_k(\tau) \tag{2.29}$$

with dimensionless time $\tau = \kappa t$, parameters $\overline{\rho} = \rho/\kappa$, $\overline{\lambda} = \lambda/\kappa$, and attachment rate $\overline{R}_k(\tau) = k/(m_0 + \overline{\lambda}\tau)$ respectively.

Let

$$\overline{H}_k(t) = \exp\left[\int \overline{R}_k(\tau)d\tau\right] = \left(m_0 + \overline{\lambda}\tau\right)^{k/\overline{\lambda}} \ , \tag{2.30}$$

so that Eq. (2.29) can be written as

$$\frac{d}{d\tau}\left[N_k(\tau)\overline{H}_k(\tau)\right] = \overline{\rho}\,\overline{H}_k(\tau)\delta_{km} + \overline{R}_{k-1}(\tau)\overline{H}_k(\tau)N_{k-1}(\tau) \ . \tag{2.31}$$

The solution of this transformed equation can be written as the following integral form

$$N_k(\tau) = \overline{\rho}\frac{(m_0+\overline{\lambda}\tau)}{k+\overline{\lambda}}\delta_{km} + \frac{(1-\delta_{km})}{\overline{H}_k(\tau)}\int \overline{R}_{k-1}(\tau)\overline{H}_k(\tau)N_{k-1}(\tau)d\tau + C_k \ , \tag{2.32}$$

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

$$N_{m+k}(\tau) = \overline{\rho}\frac{(m)_k}{(m+\overline{\lambda})_{k+1}}\left(m_0+\overline{\lambda}\tau\right) + \sum_{i=0}^{k}\frac{(m)_k}{(m)_i}\frac{C_{m+i}}{(k-i)!}\left(m_0+\overline{\lambda}\tau\right)^{-(m+i)/\overline{\lambda}} \tag{2.33}$$

where $(\gamma)_j \equiv (\gamma)(\gamma+1)\ldots(\gamma+j-1)$ is the Pochhammer symbol. The last step towards a complete solution is to determine an explicit form of the constants of integrations $\{C_{m+k}\}$ in terms of the initial conditions $\{N_{m+k}(0)\}$. This is easily accomplished by writing (2.33) in a matrix form for the vector of initial conditions $\boldsymbol{N}(0)$

$$\boldsymbol{N}(0) = \boldsymbol{A}(0) + \boldsymbol{L}(0)\boldsymbol{C} \tag{2.34}$$

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

$$L_{m+k,m+i}(0) = \binom{m+k-1}{m+i-1}\frac{1}{m_0^{m+i}} \tag{2.35}$$

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

$$M_{m+k,m+i} = (-1)^{k-i}\binom{m+k-1}{m+i-1}m_0^{m+i} \ . \tag{2.36}$$

Inserting this solution in (2.33), we get

$$\boldsymbol{N}(\tau) = \left[\boldsymbol{A}(\tau) - \boldsymbol{L}(\tau)\boldsymbol{M}\boldsymbol{A}(0)\right] + \boldsymbol{L}(\tau)\boldsymbol{M}\boldsymbol{N}(0) \ , \tag{2.37}$$

which nicely isolates the principal dynamics (the first 2 terms) from the initial conditions. Specifically, by imposing the usual initial conditions, $N_{m+k}(0) = \delta_{k0}$, it is straightforward, albeit somewhat lengthy, to obtain a closed-form expression for the complete dynamical elements as

$$\begin{aligned} N_{m+k}(\tau) &= (m)_k\frac{1}{\Gamma(k+1)}X(\tau)^m(1-X(\tau))^k \\ &\quad + \overline{\rho}m_0(m)_k\left[\frac{1}{(m+\overline{\lambda})_{k+1}}X(\tau) - \frac{1}{(m+\overline{\lambda})}\frac{1}{\Gamma(k+1)}X(\tau)^m F_k(X(\tau))\right] \end{aligned} \tag{2.38}$$

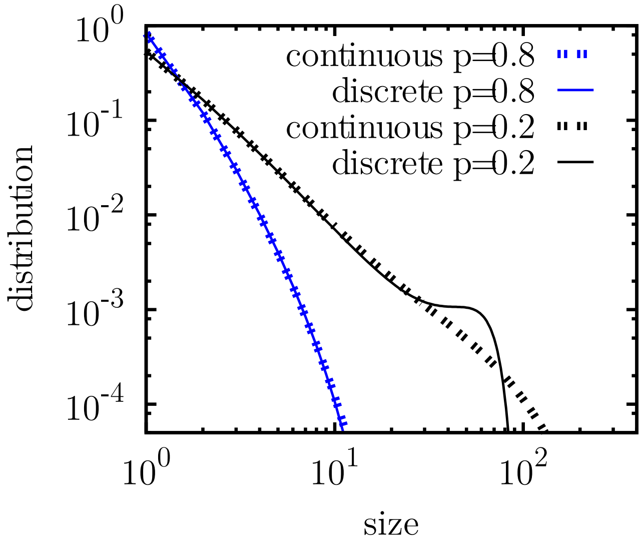


Figure 2.4 – **Discrete versus continuous preferential attachment.** Comparison of distributions obtained with $p = 0.8$ and $p = 0.2$ in discrete and continuous dynamics at time $t = 100$. This illustrates how the peloton dynamics is a direct consequence of the maximal system size present only in the discrete version of the process.

with $X(\tau) = m_0/(m_0 + \overline{\lambda}\tau)$ and where $F_k(X) = {}_2F_1(-k, m + \overline{\lambda}; m + \overline{\lambda} + 1; X)$ represents a hypergeometric series of degree k:

$$ {}_2F_1\left(a, b; c; z\right) = \sum_{n=0}^{\infty} \frac{(a)_n (b)_n}{(c)_n} \frac{z^n}{n!} \tag{2.39} $$

which is terminated when the Pochhammer symbol $(a)_n$ with $a = -k$ reaches 0. One verifies that, by setting $\tau = 0$ in the obtained solution, one obtains $N_{m+k}(0) = \delta_{k0}$ as it should.

Figure 2.4 presents a comparison of PA time evolution in discrete and continuous time. The two solutions differ for small p as the transition probabilities for elements with large k are then significant, such that the continuous approximation fails. Corrections can be applied by considering higher order terms in the expansion [5] of the discrete master equation. However, it can be shown that the continuous and discrete time versions of PA converge toward the same asymptotic behaviour.

2.2.1.2 Peloton dynamics

One particularly interesting feature of the results presented in Fig. 2.4 is the dynamics of the bins in the tail of the distributions. These bins grouped in clearly identifiable *bulges* only in the discrete version and when $p < 0.5$ such that the leading bins remain distinct from the

[5] Kramers-Moyal expansion: essentially equivalent to a Taylor series and allows to consider derivatives of higher order (i.e., a sum over $\partial^m N/\partial k^m$).

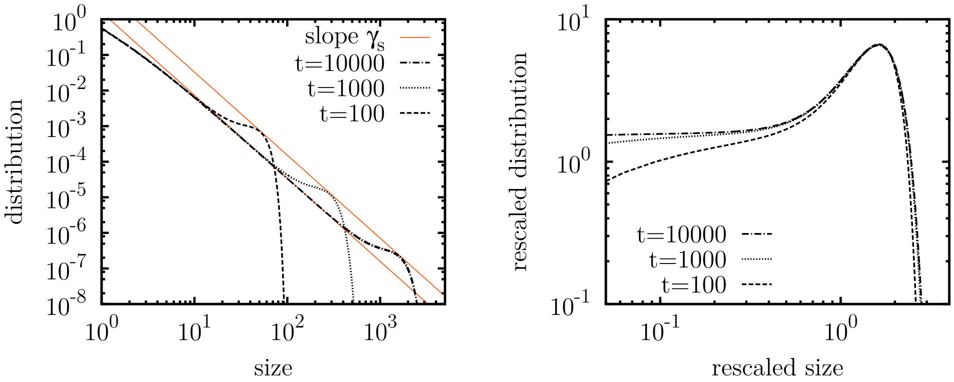


Figure 2.5 – **Rescaling of preferential attachment peloton.** (left) The height of the peloton follows a power-law decay (here for $p = 0.25$), such that its surface is conserved on a logarithmic scale as it evolves. The decay exponent of the peloton is the same as the scaling exponent of the distribution it creates. (c) Rescaled distribution $\{n^{\gamma_s} S_n(t)\}$ as a function of rescaled community size n/t^{1-p} highlights the scaling of the peloton dynamics.

new bins. The dynamics of a system's leader is well-documented in the context of growing networks [66, 50] or word frequencies [16], but can be applied to any problem where one is interested in the statistics of the extremes (i.e., the growth of the biggest business firm, of the most popular website, etc.). What we observe here is that averaging over multiple realizations of the same experiment will result in the creation of a *peloton* where one is significantly more likely to find entities than predicted by the asymptotic distribution (i.e., the leaders).

As seen on Fig. 2.4, the clear distinction between the statistical distribution of leaders versus the rest of the system is a consequence of the maximal size of the system and of the limited growth resources available. As mentioned earlier, the continuous time version of PA has no finite limitation to the number of growth events at every time step. Comparing the results of the discrete and continuous versions of our stochastic process on Fig. 2.4 illustrates how limiting growth resources results in the condensation of the leaders in a peloton. This draws a strong parallel between discrete preferential attachment and some sandpile models known to result in scale-free avalanche size distributions through *self-organized criticality*[6]. In some cases, such as the Oslo model (see [26] §3.9), the biggest avalanches are limited by the size of the considered sandpile and are thus condensed in bulges identical to our pelotons.

Also striking is the fact that this peloton conserves its shape on a log-log scale (see Fig. 2.5(left)). To highlight this feature, Fig. 2.5(right) rescales the distributions to account for the scaling in size (γ_s) and the peloton growth through time (t^{1-p}). This rescaling method was borrowed from [26] §3.9.8 and is based on the behaviour of $k_{\max}(t)$.

Using the general form of PA given above in terms of a growth rate $1 - p$, we can follow the

[6] Note that we will return, in a different context, on self-organized criticality in Chap. 4.

average position of $k_{\max}(t)$ by writing

$$k_{\max}(t+1) = \left(1 + \frac{1-p}{t}\right) k_{\max}(t) \ , \tag{2.40}$$

which directly fixes the derivative in the limit of large t,

$$\frac{d}{dt} k_{\max}(t) = \frac{1-p}{t} k_{\max}(t) \ . \tag{2.41}$$

The general solution to Eq. (2.41) is

$$k_{\max}(t) = At^{1-p} \ . \tag{2.42}$$

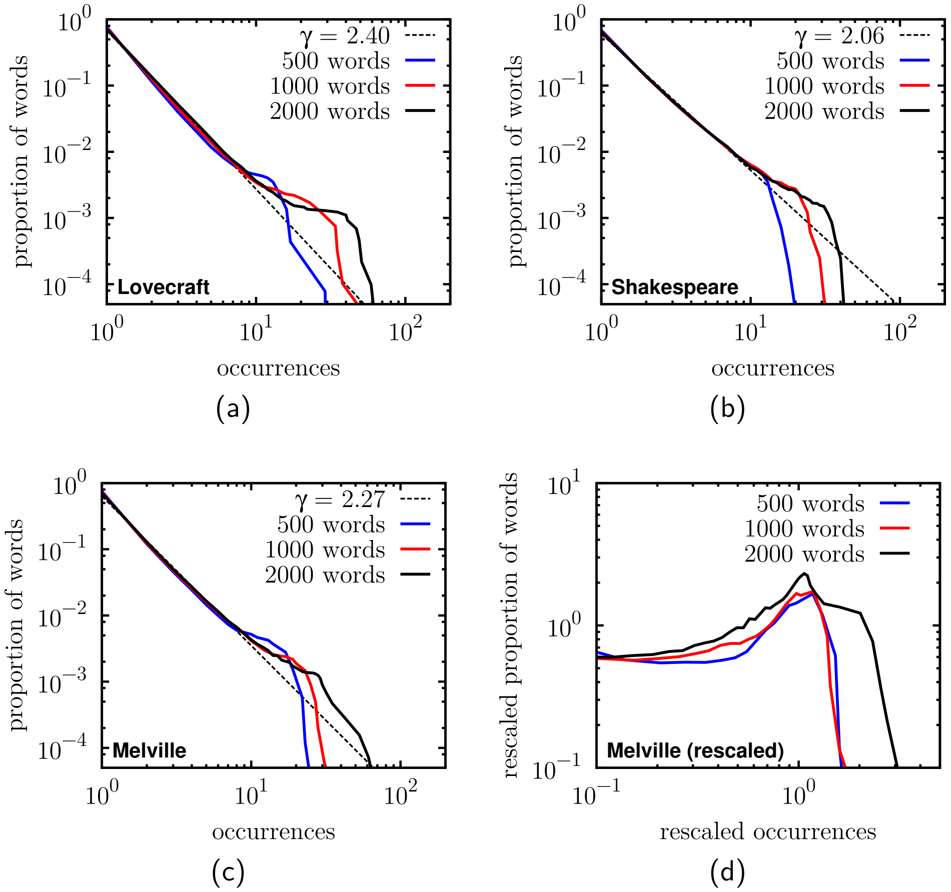
Fixing the initial condition $k_{\max}(1) = 1$, one obtains the following mean position at time t:

$$k_{\max}(t) = t^{1-p} \ . \tag{2.43}$$

This result fixes the rescaling used in Fig. 2.5(left) and also confirms that the largest bin is solely responsible for the observed peloton. Hence, different leaders might emerge in every single preferential growth realization, but the peloton dynamics can only manifest itself through averaging. This average can be taken over multiple systems or, if the dynamics include death events, over many characteristic time scales of a single system across the births and deaths of many different leaders. Consequently, empirical observations of this phenomenon are rare, because on the one hand we have only one Internet, one arXiv, and basically a unique copy of most complex systems, and on the other hand, we rarely have access to extensive data through long time scales. We can however find a solution if we go back to one of our first examples: the scale-free distribution of words by their number of occurrences in written text. Remember that in this context, the p parameter corresponds to the probability that each new written word has never been used before. We can therefore consider different samples of text of equal length written by the same author as different realizations of the same experiment.

With this in mind, we have picked different authors according to personal preferences and size of their body of work and divided their œuvres in samples of given lengths which we then used to evaluate Zipf's law under averaging (see Fig. 2.6). As predicted by PA, taking the average of multiple realizations of the same experiment results in a peloton which diverges from the traditional Zipf's law. In this case, the peloton implies that the leaders of this system (i.e., the most frequent words) consistently fall within the same range of occurrences.

Lastly, Fig. 2.6d reproduces the scaling analysis of Fig. 2.5(right) for empirical results on prose samples. The varying surface of the peloton hints at a non-constant vocabulary growth rate: a well-known feature of written text (see [54] §7.5). A first form of time-dependent birth rate [103], motivated by empirical observations, will be investigated at the end of this chapter. Yet, we now complete our study of classical preferential attachment models by looking at the impact of the attachment kernel. Hitherto, we have focussed on a *linear* relation between



(a)

(b)

(c)

(d)

Figure 2.6 – **Peloton dynamics in prose.** Distributions of words by their number of occurrences in prose samples of different length taken from the complete works of (a) H.P. Lovecraft composed of nearly 800 000 words, (b) William Shakespeare with around 900 000 words and (c) Herman Melville with over 1 200 000 words. The peloton dynamics is manifest in all distributions. (d) The rescaling method of Fig. 2.5(right), with $\gamma = 2.27$ and $1 - p = 0.43$, is applied to the statistics of Herman Melville's work.

one's current share of the resource and one's future growth, but we now consider a somewhat more general framework. This generalization of preferential attachment, along with the time-dependent results to be considered after, lead to our main contribution which will be presented in the next chapter.

2.2.2 Preferential attachment with arbitrary kernels

In this section, we investigate the impact of the attachment kernel on the result of a preferential attachment process following the work of Krapivsky and Redner [67, 65]. More generally, instead of a strictly linear kernel, where balls are thrown in bins directly preferentially to their size k, we consider a general kernel of the form $G(k) = k^{\nu}$. Consequently, we study any process where the evolution of the number of elements $N_k(t)$ with share k at time t follow a master equation of the type

$$N_k(t+\Delta t) = N_k(t) + B\delta_{k,1} + \frac{G}{\mu t}\left[(k-1)^\nu N_{k-1}(t) - k^\nu N_k(t)\right] \ , \tag{2.44}$$

where B and G are the probabilities that birth and growth events occur during the time step Δt, and where the exponent ν represents the attachment kernel. Simon's model of linear preferential attachment can then be obtained by setting $\nu = 1$, $B = p$ and $G = 1 - p$. Thus, $\nu = 1$ leads to a direct rich-get-richer process, whereas the case $\nu < 1$ represents a rich-get-richer process with diminishing returns and $\nu > 1$ the opposite. Whatever the case, μ is the average contribution of a time step to the normalization of the kernel, i.e. $\mu t = \sum_k k^\nu N_k(t)$. For example, with $\nu = 1$ we can directly calculate that $\mu = B + G$.

Since B is the probability of a birth event, the evolution of the normalized distribution $\{n_k(t)\}$ can be obtained by replacing $N_k(t)$ by $Btn_k(t)$:

$$B\left(t+\Delta t\right) n_k(t+\Delta t) = Btn_k(t) + B\delta_{k,1} + \frac{GB}{\mu}\left[(k-1)^\nu n_{k-1}(t) - k^\nu n_k(t)\right] \ . \tag{2.45}$$

Since Δt is an arbitrary time step, and B and G are just as arbitrary, we can use an equivalent process in continuous time by using $\Delta t \to dt$,

$$B\left(t+dt\right) n_k(t+dt) = Btn_k(t) + dt\left\{\frac{GB}{\mu}\left[(k-1)^\nu n_{k-1}(t) - k^\nu n_k(t)\right] + B\delta_{k,1}\right\} \ , \tag{2.46}$$

from which the following set of ordinary differential equations is obtained:

$$\lim_{dt\to 0}\frac{(t+dt)\, n_k(t+dt) - tn_k(t)}{dt} = \frac{d}{dt}\left[tn_k(t)\right] = \frac{G}{\mu}\left[(k-1)^\nu n_{k-1}(t) - k^\nu n_k(t)\right] + \delta_{k,1} \ . \tag{2.47}$$

Solving at statistical equilibrium, i.e. $n_k(t) = n_k^*$ such that $\frac{d}{dt}n_k(t) = 0 \ \forall \ k$, yields

$$\left(1 + k^\nu \frac{G}{\mu}\right) n_k^* = \frac{G}{\mu}\left(k-1\right)^\nu n_{k-1}^* + \delta_{k,1} \tag{2.48}$$

or more directly

$$n_k^* = \frac{\prod_{m=1}^{k-1}\frac{G}{\mu}m^\nu}{\prod_{m=1}^{k}\left(1 + m^\nu \frac{G}{\mu}\right)} = \frac{\mu\prod_{m=1}^{k-1} Gm^\nu}{\prod_{m=1}^{k}\left(\mu + Gm^\nu\right)} \ . \tag{2.49}$$

This last expression can be analysed through a self-coherent argument using the general definition of the normalization factor

$$\mu = \sum_k k^\nu n_k^* \ . \tag{2.50}$$

More detailed solutions of the asymptotic steady state are kernel dependent and will be investigated in what follows.

2.2.2.1 Linear connection kernel ($G(k) = k$)

In the upcoming sections, we are going to assume a simple case of network growth where $B = G = 1$. This corresponds to a process where, at each time step Δt, a new node is added to the network and connected to a single existing node according to a given kernel ν. The resource k under study is then the degree of each node after t nodes/links have been added to the system.

The simplest version of this process uses the linear kernel that has been simply assumed up to now: i.e., $\nu = 1$. In that context, Eq. (2.49) yields

$$n_k^* = \frac{2\prod_{m=1}^{k-1} m}{\prod_{m=1}^{k}(2+m)} = n_{k-1}^* \frac{k-1}{2+k} \; , \tag{2.51}$$

a telescopic product that we can reduce to

$$n_k^* = \frac{4}{k\,(k+1)\,(k+2)} \; . \tag{2.52}$$

Hence, a strictly linear kernel with $B = G = 1$ leads to $n_k^* \propto k^{-3}$; which is consistent with our previous results for the Barabási-Albert model and for the Simon's model using $p = B/(B+G) = 1/2$ such that $\gamma = (2-p)/(1-p) = 3$.

A very interesting consequence of this simple model is that we can tune the scale exponent by limiting the kernel only to an asymptotic linear behaviour. We use a general attachment probability equal to $G(k) = a_k$ (instead of k^ν) for nodes of degree k and restrict it to follow $a_k = a_\infty k$ for $k \gg 1$ only. All other probabilities are free, meaning we use arbitrary a_k for smaller k values. The general solution follows a form similar to Eq. (2.49), once again with $G = 1$,

$$n_k^* = \frac{\prod_{m=1}^{k-1} a_m/\mu}{\prod_{m=1}^{k}(1 + a_m/\mu)} = \frac{\mu}{a_k}\prod_{m=1}^{k}\left(1+\frac{\mu}{a_m}\right)^{-1} \; , \tag{2.53}$$

which has already been shown to scale as $k^{-\gamma}$ with $\gamma = 1 + (B+G)/G$ which here gives ($G = 1$ and $B + G \equiv \mu$)

$$\gamma = 1 + \mu \; . \tag{2.54}$$

This tunable exponent is thus fixed with

$$\mu = \sum a_k n_k^* = \mu \sum_{k=1}^{\infty}\prod_{m=1}^{k}\left(1+\frac{\mu}{a_m}\right)^{-1} \; . \tag{2.55}$$

This last expression implies

$$\sum_{k=1}^{\infty}\prod_{m=1}^{k}\left(1+\frac{\mu}{a_m}\right)^{-1} = 1 \tag{2.56}$$

for all $\{a_m\}$, such that

$$\left(1+\frac{\mu}{a_1}\right)^{-1}\left[1+\sum_{k=2}^{\infty}\prod_{m=2}^{k}\left(1+\frac{\mu}{a_m}\right)^{-1}\right] = 1 \tag{2.57}$$

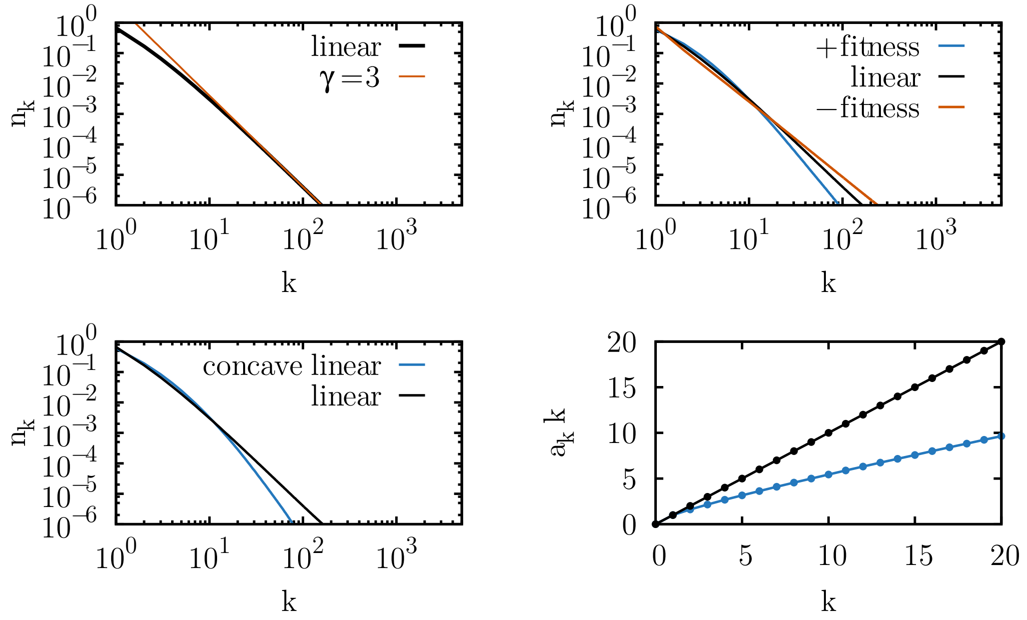


Figure 2.7 – **Results on different linear kernels.** (top left) Distribution after the introduction 50, 000 nodes using a strictly linear connection kernel. The theoretical steady-state scale exponent is shown for comparison. (top right) The same system is now compared with systems of the same size with linear kernels of the form $G(k) = a_k = k+1$ (positive initial fitness) and $G(k) = a_k = k-1/2$ (negative initial fitness). Those kernels are described in greater details in the next section. (bottom left) Result of the linear kernel is now compared with an asymptotically linear kernel, i.e. $G(k) = \left(k + 2\sqrt{k}\right)/3$. (bottom right) Kernels used for the bottom left figure. Note that all distributions are obtained by iterating Eq. (2.44) with $B = G = 1$ and slightly modified to consider the appropriate kernel.

or, as a simpler expression,

$$\mu = a_1 \sum_{k=2}^{\infty} \prod_{m=2}^{k} \left(1 + \frac{\mu}{a_m}\right)^{-1} . \tag{2.58}$$

Thus, this case is general enough to reproduce all exponents greater than two (as the case considered here) even with fixed birth and growth rates. Moreover, a_1 can then be elegantly interpreted as an initial fitness or attractiveness.

Figure 2.7 presents results of system growth based on the model considered in this section using different, strictly or asymptotically linear connection kernels.

2.2.2.2 Sub-linear connection kernel ($G(k) = k^{\nu}$ with $\nu < 1$)

We now consider the case of a rich-get-richer process with diminishing returns, i.e. $\nu < 1$. Rewriting the steady state result of the general process gives

$$n_k^* = \frac{\prod_{m=1}^{k-1} m^{\nu}/\mu}{\prod_{m=1}^{k} (1 + m^{\nu}/\mu)} = \frac{\mu}{k^{\nu}} \prod_{m=1}^{k} \left(1 + \frac{\mu}{m^{\nu}}\right)^{-1} . \tag{2.59}$$

To solve the behaviour of n_k^* for large k, we write

$$n_k^* = \frac{\mu}{k^\nu}\exp\left[-\sum_{m=1}^{k}\ln\left(1+\frac{\mu}{m^\nu}\right)\right] \sim \frac{\mu}{k^\nu}\exp\left[-\int_1^k \ln\left(1+\frac{\mu}{m^\nu}\right)dm\right] . \tag{2.60}$$

The general solution is given in terms of the hypergeometric function ${}_2F_1\left(a,b;c;z\right)$ [1],

$$n_k^* \sim \frac{\mu}{k^\nu}\exp\left[\nu k \cdot {}_2F_1\left(1,-1/\nu,1-1/\nu,-\mu k^{-\nu}\right) - k\ln\left(\mu k^{-\nu}+1\right) - \nu k + \text{constant}\right] . \tag{2.61}$$

We now keep only the non-vanishing functions of k in the exponential. To do so, we expand every function in its power series. Let us recall that the hypergeometric series takes the form

$${}_2F_1\left(a,b;c;z\right) = \sum_{n=0}^{\infty}\frac{(a)_n(b)_n}{(c)_n}\frac{z^n}{n!} \tag{2.62}$$

where $(a)_n$ is again the Pochhammer symbol. The first terms thus removes the linear function of k and we combine the remaining terms with the power series of the logarithm:

$$n_k^* \sim \frac{\mu}{k^\nu}\exp\left[\nu\sum_{n=1}^{\infty}(-1)^n\left\{\frac{(1)_n(-1/\nu)_n}{n!(1-1/\nu)_n}+\frac{1}{n\nu}\right\}\mu^n k^{1-n\nu} + \text{constant}\right] . \tag{2.63}$$

The important point here is that keeping only the positive powers of k in the exponential imply keeping all powers in the series such that $1-n\nu>0$. Meaning that if $\nu>1/2$, we keep only the first term; or more generally if for a natural number m we find $1/(m+1)<\nu<1/m$, we keep only the first m terms in the series. The special case where $\nu=1/m$ must be accounted for separately as the hypergeometric falls back on a logarithmic function. For instance, if $\nu=1/2$, the integral becomes

$$n_k^* \sim \frac{\mu}{\sqrt{k}}\exp\left[-\int_1^k \ln\left(1+\frac{\mu}{\sqrt{m}}\right)dm\right] \tag{2.64}$$

$$\sim \frac{\mu}{\sqrt{k}}\exp\left[2\mu^2\ln\left(\mu+\sqrt{k}\right) - k\ln\left(1+\frac{\mu}{\sqrt{k}}\right) - \mu\sqrt{k} + \text{constant}\right] , \tag{2.65}$$

$$\tag{2.66}$$

where the last two functions in the exponential can be treated as before, but the first term modifies the power of k outside of the exponential (assuming $\sqrt{k}\gg\mu$ since $1<\mu<2$ for $0<\nu<1$). To summarize, the first interval $1/3<\nu<1$ is given by

$$n_k^* \sim \begin{cases} k^{-\nu}\exp\left[-\mu\frac{k^{1-\nu}}{1-\nu}\right] & \text{if } 1/2<\nu<1 \\ k^{(\mu^2-1/2)}\exp\left[-2\mu\sqrt{k}\right] & \text{if } \nu=1/2 \\ k^{-\nu}\exp\left[-\mu\frac{k^{1-\nu}}{1-\nu}+\frac{\mu^2}{2}\frac{k^{1-2\nu}}{1-2\nu}\right] & \text{if } 1/3<\nu<1/2\,. \end{cases} \tag{2.67}$$

These distributions all behave as stretched exponential (or power-law with strong cut-offs) distribution. Finally, the normalization is once again obtained using Eq. (2.58) with, in this case, $a_1=1$.

2.2.2.3 Supra-linear connection kernel ($G(k) = k^\nu$ with $\nu > 1$)

A preferential attachment process with a supra-linear kernel leads to a rich-get-richer process with increasing returns: meaning it is progressively easier for the rich to get richer. Simple arguments can be shown to illustrate how this leads to a condensate phase where most, if not all, links are attached to a single node. For instance, if the leading node is currently connected to all other $N-1$ nodes, it will connect to the next introduced node with probability $(N-1)^\nu/[N-1+(N-1)^\nu]$. Starting from the beginning, this leading node will maintain its state indefinitely with probability

$$\mathcal{P} = \prod_{N=0}^{\infty} \frac{1}{1+N^{1-\nu}} \simeq \exp\left[-\int_0^\infty \ln\left(1+N\right)^{1-\nu} dN\right] \simeq \exp\left[-\int_0^\infty N^{1-\nu} dN\right] \tag{2.68}$$

which is a non-zero probability for $\nu > 2$. For a more general description, we can directly follow the $N_k(t=k)$ as the probability of having a fully connected condensate (assuming the leading node is present as an initial condition at time $t=0$). Rewriting Eq. (2.44) specifically for this state, using $N_k(t) = 0 \ \forall k > t$, yields

$$N_k(k) = \frac{(k-1)^\nu N_{k-1}(k-1)}{K(k-1)} \tag{2.69}$$

where $K(k-1)$ is the time dependent normalization factor (previously μt in the long-time limit). Simplifying the description using the fact that $N_2(2) = 1$, as there must be a node with degree 2 when only three nodes are in the system, leads to

$$N_k(k) = \prod_{k'=2}^{k-1} \frac{(k')^\nu}{K(k')} \tag{2.70}$$

with $K(t)$ constrained by the lower bound t^ν, as no terms in the product can be higher than one, and by the higher bound

$$K(t) = \sum_{k=1}^{t} k^\nu N_k(t) \leq t^{\nu-1} \sum_{k=1}^{t} k N_k(t) \sim t^\nu \tag{2.71}$$

as we know the normalization factor scales linearly with time for $\nu = 1$ regardless of the distribution at time t. Thus, $K(t) \sim t^\nu$. As $K(t)$ is not strictly equal to t^ν, we can not directly insert it in Eq. (2.70) and must instead iteratively solve $N_k(t)$. We first concentrate on the leading behaviour (i.e., leading power in t) for each $N_k(t)$. First for $N_1(t)$, we write

$$\dot{N}_1(t) = 1 - \frac{N_1(t)}{t^\nu} \;\Rightarrow\; N_1(t) \sim t \tag{2.72}$$

as the first term in the rate equation governs the leading behaviour. Similarly for $N_2(t)$, we find

$$\dot{N}_2(t) = \frac{N_1(t) - 2^\nu N_2(t)}{t\nu} = t^{1-\nu} - \frac{2^\nu}{t^\nu} N_2(t) \;\Rightarrow\; N_2(t) \sim \frac{t^{2-\nu}}{2-\nu} \, . \tag{2.73}$$

Continuing this simple integration of the leading terms in each rate equation uncovers the following pattern,

$$N_k(t) \sim \left[\prod_{j=1}^{k-1} \frac{j^\nu}{1 + j\,(1-\nu)}\right] t^{k-(k-1)\nu} \ . \tag{2.74}$$

The first correction term can then be obtained by reintroducing $N_1(t)$ within the rate equations and now keeping two terms; for instance $N_1(t) \sim t - t^{2-\nu}/\,(2-\nu)$. Further correction terms could then be obtained by reapplying the procedure, i.e. $N_1(t) \sim t - t^{2-\nu}/\,(2-\nu) + t^{3-2\nu}/\,[(2-\nu)\,(3-2\nu)] - \ldots$

However, we now focus on the leading behaviour to highlight an interesting property. The leading power of t given by Eq. (2.74) implies that $N_k(t)$ becomes finite once $k-(k-1)\nu \leq 0$, which occurs for degree k such that

$$(k-1)\,\nu \geq k \ \Rightarrow \ k \geq \frac{\nu}{\nu - 1} \ . \tag{2.75}$$

The key result here is that for $(m+1)/m < \nu < m/(m-1)$ we have a finite number of nodes with more than m links. These nodes thus constitute a condensate of the whole system as a finite number of nodes will end up sharing an infinite number of links with infinitely many new nodes. The finiteness of all $N_k(t)$ beyond a certain degree also implies that the share $k_{\max}$ of the leading node considered earlier, i.e. node $N_k(k)$ or $N_{k_{\max}}(t)$, must scale linearly with time (since the number of introduced links equals t).

2.2.2.4 Additional results and discussions on attachment kernels

The results of this section perfectly illustrate how the power-law organization is a critical state of resource distribution; or, equivalently, how (linear) preferential attachment is a critical growth process. As mentioned in the introduction, criticality refers to a property of systems at phase transition. In the context of resource distribution, the two different phases (or states) at play are a homogeneous or fair system on the one hand, and a condensate or unfair winner-takes-all scenario on the other hand. Results obtained on linear and sub- or supra-linear kernels are presented in Figure 2.8 to highlight the difference between phases and their link with the different kernels.

A slightly different way of visualizing these different cases as different phases is to study the moments of their respective resource distribution in the long-time or large-size limit ($t \to \infty$ or $N \to \infty$). In the homogeneous case, all moments are well-defined. Whereas all moments but the first (the average share $\langle k \rangle = (B+G)/B$) will diverge in the condensate case. The critical state is thus characterized by a tunable number of diverging moments (following the tunable scale exponent).

The non-universality of the scaling exponent, i.e. its dependence on the model parameters, is an essential feature of a general preferential attachment process as different scale-independent

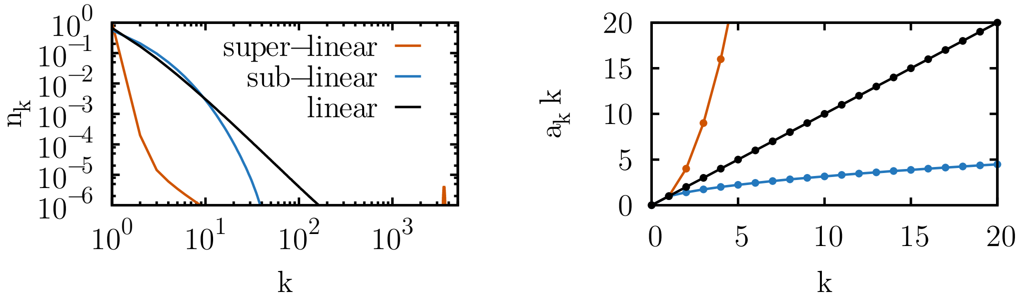


Figure 2.8 – **Different kernels and phases for resource distribution.** (left) Distribution obtained using strictly linear, sub- and supra-linear connection kernels after the introduction of 50,000 nodes. Note the condensate, or leading node, at around k =3,500 in the supra-linear case. Results are obtained by iterating Eq. (2.44) with $B = G = 1$ and the appropriate kernel. (right) Kernels used for the left figure: $G(k)$ equal to k, $\sqrt{k}$ and k^2 for the linear, sub- and supra-linear cases respectively (or $\nu = 1$, 1/2 and 2).

systems often show different scaling exponents. Modifying the connection kernel is one of two general mechanism that allows us to tune the exponent to the desired value. The second involves time-dependent birth and growth probabilities and will be studied in the next section. For the moment, suffice it to say that the connection kernel can itself model more complicated generalizations of the classic preferential attachment process. For instance, the ageing of sites, and its consequent effect on sites' attractiveness, has also been shown to affect and potentially break the scaling behaviour [33]. Yet, share-age correlations and age-attractiveness correlations can always be collapsed into a coherent kernel modification (or share-attractiveness correlations); especially considering that nodes of equal shares are usually considered as indistinguishable.

2.2.3 Preferential attachment with time-dependent birth rate

In this section, we investigate the effect of a time-dependent birth rate [103]. To this end, we will use Heaps' law, first hinted at in Sec. 2.2.1.2, which states $p(t) \propto t^{-\alpha}$ with $\alpha \in [0, 1]$. Using a linear preferential attachment and $p(t) = t^{-\alpha}$, we now wish to determine the qualitative steady-state behaviour as a function of α. Rewriting the master equation, we get

$$N_1(t+1) = N_1(t) - \frac{1-t^{-\alpha}}{t}N_1(t) \tag{2.76}$$

$$N_k(t+1) = N_k(t) + \frac{1-t^{-\alpha}}{t}\left[(k-1)N_{k-1}(t) - kN_k(t)\right] \qquad \text{for } k > 1. \tag{2.77}$$

As we are only interested in the scale exponent, we can get a rough estimate by approximating both the time and the distribution as continuous. Basically, instead of looking for the ensemble of solutions $\{N_k(t)\}$, we will look for an approximated continuous probability density $P(x, t)$

(i.e. $P(x,t) \sim N_x(t)$). Equations (2.76) and (2.77) become

$$P(1,t+dt) = P(1,t) + dt\left[t^{-\alpha} - \frac{1-t^{-\alpha}}{t}P(1,t)\right] \tag{2.78}$$

$$P(x,t+dt) = P(x,t) + dt\left[\frac{1-t^{-\alpha}}{t}\frac{(x-dx)P(x-dx,t) - xP(x,t)}{dx}\right] \tag{2.79}$$

equivalent to the following differential equations

$$\frac{\partial}{\partial t}P(1,t) = t^{-\alpha} - \frac{1-t^{-\alpha}}{t}P(1,t) \tag{2.80}$$

$$\frac{\partial}{\partial t}P(x,t) + \frac{1-t^{-\alpha}}{t}\frac{\partial}{\partial x}\left[xP(x,t)\right] = 0\ . \tag{2.81}$$

We are interested in analysing the behaviour in x for $t \gg 1$. The solution for $P(1,t)$ is readily obtained if we keep only the leading temporal term, i.e.

$$P(1,t) \simeq At^{1-\alpha} \tag{2.82}$$

where A is an arbitrary constant. The partial differential equation for the rest of the solution is more problematic. As we are looking for a qualitative description of the spatial behaviour, the methods of characteristics can be extremely useful here. We first rewrite it as

$$\frac{\partial}{\partial t}\left[xP(x,t)\right] + x\frac{1-t^{-\alpha}}{t}\frac{\partial}{\partial x}\left[xP(x,t)\right] = 0\ . \tag{2.83}$$

We now look for a parametrization of our function as $Q(x(s),t(s)) = x(s)P(x(s),t(s))$, such that Eq. (2.83) is actually equivalent

$$\frac{d}{ds}Q(x(s),t(s)) = \frac{\partial Q}{\partial t}\frac{dt}{ds} + \frac{\partial Q}{\partial x}\frac{dx}{ds} = 0\ . \tag{2.84}$$

Comparing Eqs. (2.83) and (2.84) yields a characteristic systems of ODEs. We first get

$$\frac{dt}{ds} = 1 \quad \Rightarrow \quad t = s \tag{2.85}$$

setting $t(1) = 1$ as an initial condition ($t = 0$ is undefined). This in turn gives

$$\frac{dx}{ds} = x\frac{1-s^{-\alpha}}{s} \quad \Rightarrow \quad x = x(1)t \cdot \exp\left[\left(t^{-\alpha} - 1\right)/\alpha\right]\ . \tag{2.86}$$

The last ODE is

$$\frac{dQ}{ds} = 0\ , \tag{2.87}$$

with initial condition $Q(x(1),1) = \Psi(x(1))$ where Ψ is an unknown function. This provides a general parametrization of the unknown solution if we solve for $x(1)$ in Eq. (2.86):

$$Q(x(t),t) = \Psi(x(1)) = \Psi\left(\frac{x}{t}\exp\left[-t^{-\alpha}/\alpha\right]\right)\ . \tag{2.88}$$

In terms of our original probability density function, we have

$$P(x,t) = \frac{1}{x}\Psi\left(\frac{x}{t}\exp\left[-t^{-\alpha}/\alpha\right]\right) \ . \tag{2.89}$$

which becomes for large t

$$P(x,t) \sim \frac{1}{x}\Psi\left(\frac{x}{t}\right) \ . \tag{2.90}$$

Inspection of the solution Eq. (2.82) for $P(1,t)$ provides the form of Ψ, so that we finally get

$$P(x,t) \propto x^{\alpha-2}t^{1-\alpha} \quad \Rightarrow \quad N_k(t) \propto k^{-\gamma} \text{ with } \gamma = 2-\alpha \in [1,2]. \tag{2.91}$$

In the next chapter, we will determine that within our general framework a scaling exponent between 1 and 2 directly implies some sort of temporal scaling in the birth rate. Moreover, we will introduce a coupling between the time-dependent birth rate and possible non-linear deviations in the attachment kernel.

Chapter 3

On growth I: Universal growth constraints of human systems

Résumé

L'indépendance d'échelle est une propriété universelle des systèmes complexes qui implique une organisation extrêmement inhomogène. La recherche s'attarde depuis longtemps à expliquer *pourquoi* des systèmes aussi variés que l'évolution, les interactions entre protéines, ou les actions en bourse, posséderaient tous cette indépendance d'échelle. Nous prenons plutôt le chemin inverse: nous supposons la présence de ce comportement et visons à expliquer *comment* il émerge, en contraste avec les modèles simplifiés jusqu'ici considérés.

Dans ce chapitre, nous montrons qu'un système dont la distribution de ressource croît vers l'indépendance d'échelle est assujetti à des contraintes temporelles strictes: la première étant l'attachement préférentiel et la seconde une nouvelle forme générale de comportement d'échelle temporel à délai. Ce délai agit comme un couplage entre les deux contraintes, ou plus précisément, entre la croissance de la population totale et la croissance des ressources d'un individu. Ces contraintes forment des trajectoires temporelles si précises que même une image instantanée d'une distribution est suffisante pour reconstruire son passé et prédire son futur. Nous validons notre approche sur plusieurs sphères d'activités humaines, de la productivité scientifique et artistique aux relations sexuelles et activités en ligne.

Summary

Scale independence is a ubiquitous feature of complex systems which implies a highly skewed organization with no characteristic scale. Research has long focused on *why* systems as varied as protein networks, evolution, and stock actions all feature scale independence. Assuming that they simply do, we focus here on describing exactly *how* this behaviour emerges, in contrast with the more idealized models usually considered.

In this chapter, we show that growing towards scale independence implies strict constraints: the first is the well-known preferential attachment principle and the second is a new general form of delayed temporal scaling. The delay acts as a coupling between the two constraints or, equivalently, between population growth and individual activity. These constraints pave a precise evolution path, such that even an instantaneous snapshot of a distribution is enough to reconstruct the past of the system and predict its future. We validate our approach on diverse spheres of human activities ranging from scientific and artistic productivity, to sexual relations, and online traffic.

3.1 Introduction

Human systems are often characterized by *extreme inequalities*. One may think of the distribution of wealth between individuals, the sizes of cities, or the frequencies of sexual activities to name a few [96, 104, 24, 78, 18]. Interestingly, inequality often tends to manifest itself through a *scale-independent* behaviour [96, 104, 24, 78, 18, 102, 92, 12, 56, 11, 10, 23]. In layman's terms, these systems are said to be scale-independent because of the absence of a characteristic scale. Taking the distribution of wealth as an example, the worldwide average income is meaningless because the variance is too wide. Neither the very poor nor the very wealthy can be reduced to average individuals; the former are too numerous while the latter are absurdly richer than the average.

Mathematically, this behaviour takes the form of a *power-law distribution*. That is, the number N_k of individuals having a share k (e.g. personal income or sexual partners) of the total resource K (total wealth or sexual activities) roughly follows $N_k \propto k^{-\gamma}$. One of the first robust observations of scale-independent systems concerns the distribution of occurrences of individual words in prose [104] as illustrated in Fig. 3.1(left).

In this chapter, we build upon two general premisses to describe the growth of scale-independent systems. Firstly, we assume that the underlying distribution roughly follows $N_k \propto k^{-\gamma}$ such that a power law is an adequate approximation for all k (with $\gamma > 1$ for normalization in the asymptotic limit). Secondly, we follow the distribution of a resource or property that can only

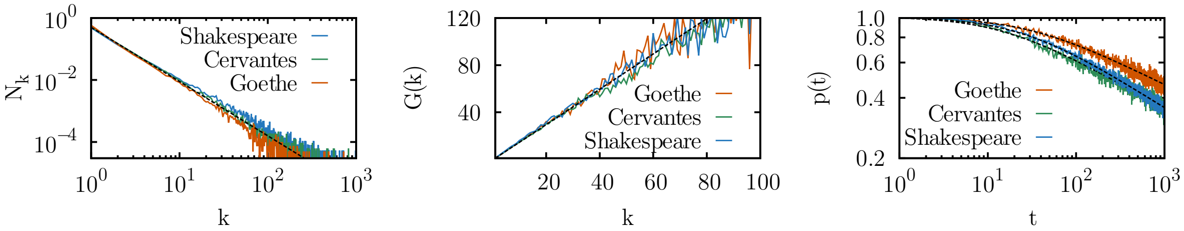


Figure 3.1 – **Scale independence, preferential attachment and delayed temporal scaling in prose samples.** (left) Power-law distribution of word occurrences in the writings of authors in three different languages. Power law with scale factor $\gamma = 1.75$ is plotted to guide the eye. Actual scale exponents are estimated[28] at 1.89 for Goethe, 1.76 for Cervantes, and 1.67 for Shakespeare. (middle) Preferential attachment in written text with a linear relation for comparison. The algorithm to obtain $G(k)$ is given in Sec. 3.4. (right) Average birth function for samples of 1000 words, this procedure is based on the translational invariance [17] of written texts and yields better statistics. Fits of Eq. (3.17) are overlaid using $[\alpha, \tau, b]$ equal to $[0.22, 31, 0]$, $[0.25, 15, 0]$ and $[0.28, 25, 0]$ (with a fixed by $p(1) = 1$)for Goethe's, Cervantes' and Shakespeare's writings respectively. This asymptotic scaling is related to what is generally known as Heaps' law of vocabulary growth in linguistics [54], but is given here a much more general expression for all t.

increase or stagnate, namely the total activities of an individual (both past and present).

We use diverse databases to validate our approach: scientific productivity of authors on the arXiv e-print archive (arXiv), one month of user activities on the Digg social news website (Digg) [70], productivity of actors on the Internet Movie Database (IMDb), sexual relations in a Brazilian escort community (sexual) [30] and the writings of William Shakespeare, Miguel de Cervantes Saavedra and Johann Wolfgang von Goethe.

3.2 Results

Let us consider the growth of a hypothetical system where each individual i possesses a share $k_i(t)$ of the total resource $K(t)$ at time t. Because the system is constantly growing, both in terms of its total population $N(t)$ and of each individual's share, time can be measured as the total number of events. These events can take one of two forms: *birth events* which increase the total population $N(t+1) = N(t) + 1$ by adding a new individual j with $k_j(t) = 1$; and *growth events* which imply $k_i(t+1) = k_i(t) + 1$ for a given individual i.

We then introduce two functions: a *birth function* $p(t)$ that prescribes the probability that the t-th event is a birth event, and a *growth function* $G(k)$ that describes the average chances (unnormalized probability) for an individual with current share k of being involved in the next growth event. Assuming that individuals with the same share are indiscernible, the average share k_i of an individual i can be followed through a mean-field model:

$$k_i(t+1) = k_i(t) + [1 - p(t)] \frac{G(k_i(t))}{\sum_j G(k_j(t))} \tag{3.1}$$

Consequently, the probability that a growth event involves *any* individual of current share k is given by $N_k(t)G(k)/\sum_{k'} N_{k'}(t)G(k')$ where $N_k(t)$ is the number of individuals with share k at time t. This yields the following master equation (for $k \in \mathbb{N}$):

$$N_k(t+1) = N_k(t) + p(t)\delta_{k,1} + [1-p(t)] \frac{N_{k-1}(t)G(k-1) - N_k(t)G(k)}{\sum_m N_m(t)G(m)} \tag{3.2}$$

with $N_0(t) = 0 \; \forall t$. For this model to be of any use, at least partial knowledge of $G(k)$ and $p(t)$ is required. Setting $G(k) = k$ and a constant $p(t)$, we retrieve the *classic linear preferential attachment* process [92]. However, our goal is to investigate the constraints imposed by the scale independence, $N_k(t) \propto k^{-\gamma}$, on the functional forms of both $p(t)$ and $G(k)$ as well as the coupling between the two.

The next two sub-sections are more technical in scope, but necessary to delineate the functional forms that will constitute the basis of the following study. Although our analysis is based on asymptotic arguments, and therefore approximate, we will demonstrate that the following expression,

$$p(t) = a(t+\tau)^{-\alpha} + b \tag{3.3}$$

together with $G(k) \sim k$ and the model of Eq. (3.2), captures the essence of the growth of diverse human activities. The form of $G(k) \propto k$, at least for k greater than a certain bound k^*, is not new, but emerges naturally from our premisses. As we will see shortly, the temporal dependence of $p(t)$ is inherent to the growth towards scale independence and is coupled to the behaviour of $G(k)$ at small k through the parameter τ.

3.2.1 The growth function

The behaviour of the growth function $G(k)$ can be constrained by an argument presented by Eriksen and Hörnquist [42]. We wish to obtain $G(k)$ solely on the basis of Eq. (3.2). Instead of measuring $G(k)$ directly by looking at what leaves $N_k(t)$, we can equivalently look at what arrives in the states $k' > k$ during the time step $t \to t+1$. We write this as the difference between what is in $k' > k$ at $t+1$ [i.e. $\sum_{i=k+1}^{\infty} N_i(t+1)$] and what was in $k' > k$ at time t [i.e. $\sum_{i=k+1}^{\infty} N_i(t)$]. We substitute $N_i(t+1)$ with Eq. (3.2) and sum over all $k' > k$:

$$\begin{aligned}\sum_{i=k+1}^{\infty} \left[N_i(t+1) - N_i(t)\right] &= \sum_{i=k+1}^{\infty} \left\{ p(t)\delta_{i,1} + [1-p(t)] \frac{N_{i-1}(t)G(i-1) - N_i(t)G(i)}{\sum_m N_m(t)G(m)} \right\} \\ &= [1-p(t)] \frac{N_k(t)G(k)}{\sum_m N_m(t)G(m)} \; . \end{aligned} \tag{3.4}$$

This last expression can be interpreted as two measures of the activity in compartment $N_k(t)$ between t and $t+1$. The left-hand side measures the mean number of arrivals in compartment $N_{k'}(t)$ with $k' > k$; i.e. the mean number of individuals which left compartment $N_k(t)$. The right-hand side is explicitly the ratio of the activity involving the k-th compartment,

$N_k(t)G(k)$, to the total growth activity, $\sum_m N_m(t)G(m)$, times the probability, $1-p(t)$, that a growth event has occurred during the time step. From this equivalence, $G(k)$ is readily obtained from Eq. (3.4):

$$G(k) = \frac{\sum_m N_m(t)G(m)}{1-p(t)} \frac{1}{N_k(t)} \sum_{i=k+1}^{\infty} \left[N_i(t+1) - N_i(t)\right] \; . \tag{3.5}$$

For $k \gg 1$, we can replace the sum by an integral, and using our only hypothesis, i.e. $N_k(t) = A(t)k^{-\gamma}N(t)$, where $A(t)$ is a normalization factor, we find:

$$G(k) \simeq \frac{\sum_m N_m(t)G(m)}{1-p(t)} \left[\frac{A(t+1)N(t+1) - A(t)N(t)}{A(t)N(t)}\right] \frac{k}{\gamma - 1} \; . \tag{3.6}$$

All factors independent of k are of no concern, since $G(k)$ only makes sense when comparing the relative values for different k. Hence, at any given time t, we finally obtain:

$$G(k) \propto k \tag{3.7}$$

at least for values of k higher than an appropriate lower bound. This linear relation between the probability of growth of an individual and its present size, *preferential attachment*, is a recurrent feature in scale-independent growth models [102, 47, 24, 92, 12, 56]. This simple derivation states once again that a scale-independent growing system implies a linear preferential attachment. See Fig. 3.1(middle) for examples.

In recent years, the idealized preferential attachment process, using $G(k) = k$ and $p(t) = p$, has been analysed to great lengths. Most studies have been concerned with the application of this process to network growth [34, 13] and have focused on solving the resulting network structure [67, 36], describing the statistics of leading nodes [66], finite-size effects [9], and its relation to other properties of complex networks such as their modular and self-similar nature [57].

3.2.2 The birth function

A time-varying birth rate $p(t)$ has been considered before, either in *ad hoc* manner [92, 103] or in a specific context [46] based on empirical observations in, for example, written texts [54] or human mobility [22]. Instead of investigating how a given $p(t)$ might influence the distribution of resource in the system, we investigate how a given distribution of resource informs us on the actual $p(t)$ of that system. In doing so, the hope is to provide a more general framework for understanding how and why scale-independent organization implies scale-independent growth.

In our model, the birth function has two important roles. First, it is equivalent to the time derivative $\dot{N}(t)$ of the population $N(t)$; and second, it constrains the growth of the largest share $k_{\max}(t)$. Two relations can be called upon to connect $N(t)$ and $k_{\max}$, and obtain a consistent functional form for $p(t)$.

The first relation is the extremal criterion [67]: $\int_{k_{\max}(t)}^{\infty} N_k(t)dk \sim 1$, intuitively meaning that the number of individuals with a maximal share is of order one. To simplify the analysis, we will assume that $k_{\max}(t) \gg 1$, such that the normalization $A(t) = \left[\sum_1^{k_{\max}(t)} k^{-\gamma}\right]^{-1}$ has converged to a constant A^*. We thus use $N_k(t) = A^* N(t) k^{-\gamma}$ in the extremal criterion and solve for $N(t)$:

$$N(t) \sim \frac{\gamma - 1}{A^*} k_{\max}^{\gamma-1}(t) \quad \rightarrow \quad \frac{N(t)}{\dot{N}(t)} = \frac{k_{\max}(t)}{(\gamma-1)\,\dot{k}_{\max}(t)} \; . \tag{3.8}$$

Note that keeping the temporal dependence of $A(t)$ yields the same result for the leading temporal term. The second important relation stems from our definition of time t (in number of events or resource K) such that $\dot{K}(t) = 1$. We write

$$\dot{K}(t) = \frac{d}{dt}\sum_{m=1}^{k_{\max}(t)} mN_m(t) = \frac{d}{dt}\left[\sum_{m=1}^{k^*} mN_m(t) + \int_{k^*}^{k_{\max}(t)} mN_m(t)dm\right] = 1 \tag{3.9}$$

where k^* is an appropriate bound for the integral approximation of the sum. Again, using $N_k(t) = A^* N(t) k^{-\gamma}$, we obtain

$$A^*\dot{N}(t)\left[C + \frac{1}{2-\gamma}k_{\max}^{2-\gamma}(t) + \frac{N(t)}{\dot{N}(t)}k_{\max}^{1-\gamma}(t)\dot{k}_{\max}(t)\right] = 1 \; , \tag{3.10}$$

where C is a constant collecting all terms independent of t. Replacing $N(t)/\dot{N}(t)$ with Eq. (3.8) allows us to solve for $\dot{N}(t)$ [i.e. $p(t)$]:

$$p(t) = \dot{N}(t) = \frac{(2-\gamma)(\gamma-1)}{A^*}\frac{1}{C(2-\gamma)(\gamma-1) + k_{\max}^{2-\gamma}(t)} \tag{3.11}$$

If $\gamma \in (1,2)$, $k_{\max}^{2-\gamma}(t)$ is the leading term and $p(t)$ decreases as $k_{\max}^{\gamma-2}(t)$; if $\gamma > 2$, $k_{\max}^{2-\gamma}(t)$ becomes negligible and $p(t)$ is essentially governed by the first two terms of the ensuing geometric series. We can summarize these results, obtained only by assuming $N_k(t) \propto k^{-\gamma}$ and $k_{\max}(t) \gg 1$, under a general form:

$$p(t) \propto \begin{cases} k_{\max}^{\gamma-2}(t) & \text{if } 1 < \gamma < 2 \\ k_{\max}^{2-\gamma}(t) + \text{constant} & \text{if } \gamma > 2 \; . \end{cases} \tag{3.12}$$

The remaining step is to establish the time dependence of $k_{\max}(t)$ to obtain the explicit temporal form of $p(t)$. In line with our asymptotic arguments, as $k_{\max}(t)$ increases beyond an appropriate bound k^*, Eq. (3.1) can be rewritten as

$$k_{\max}(t+1) = \left[1 + \frac{1 - p(t)}{\kappa\,(t+\tau)}\right] k_{\max}(t) \; . \tag{3.13}$$

The denominator represents the asymptotic behaviour of the normalisation of growth probabilities. One can in fact verify that the sum converges to a linear function of time such that $G(k)/\sum_k G(k)N_k(t) = [\kappa(t+\tau)]^{-1}$ for $t \gg 1$, irrespectively of the initial behaviour of $G(k)$. This initial behaviour can however offset the value of the sum by a constant, encapsulated in the factor $\kappa\tau$.

Equation (3.13) determines the derivative in the limit of large t:

$$\frac{d}{dt}k_{\max}(t) = \frac{1-p(t)}{\kappa\,(t+\tau)}k_{\max}(t)\ . \tag{3.14}$$

Since $p(t)$ is limited to the range $[0,1]$ we can write, without loss of generality, $p(t) = f(t) + b$ where b is the asymptotic value of $p(t)$. This form yields the exact solution:

$$k_{\max}(t) = C_1(t+\tau)^{(1-b)/\kappa}\mathrm{exp}\left[-\int_{t^*}^{t}\frac{f(t')}{\kappa\,(t'+\tau)}dt'\right] \tag{3.15}$$

where t^* is an appropriate lower bound such that Eq. (3.14) is applicable. As $f(t)$ is bounded, the exponential factor converges rapidly to one and we find the general solution for large t:

$$k_{\max}(t) = C_1(t+\tau)^{(1-b)/\kappa}\ . \tag{3.16}$$

Inserting Eq. (3.16) in Eq. (3.12), we obtain a functional form for the birth function:

$$p(t) \simeq a\,(t+\tau)^{-\alpha} + b\ , \tag{3.17}$$

where we identify $\alpha = (2-\gamma)/\kappa$ for $1 < \gamma < 2$ and $\alpha = (\gamma-2)\,(1-b)\,/\kappa$ for $\gamma > 2$. The first confrontation of Eq. (3.17) with empirical data is displayed in Fig. 3.1(right).

Before we describe in the next sub-section the procedure adopted to optimise the parameters $[\alpha, \tau, b]$ (the parameter a is fixed by population size) on actual data, a few comments appear necessary. These three free parameters *do not* overparameterize the function. Two of them, α and b, govern the scale exponent in the two fundamentally different regimes $\gamma < 2$ and $\gamma > 2$ respectively, while the delay τ embodies an intrinsic coupling between population growth and individual growth. For instance, a large value of τ expresses the fact that the system features strong diminishing returns on growth for small k (concave $G(k)$). To a lesser extent, κ plays a similar role, but combined with other temporal (b) and organizational (γ) features within α. Moreover, from the asymptotic nature of our derivation, it is not to be expected that the relations between the exponents α and γ should be strictly observed. However, the results of Fig. 3.1 (see the numerical values in the caption) indicates that it is nearly true for the three prose samples studied (cases with $b = 0$) : the observed $2-\kappa\alpha = 1.80, 1.76, 1.72$ are close to the inferred estimates of $\gamma = 1.89(4), 1.76(3), 1.67(8)$ respectively. For the cases where $b \neq 0$, the classical preferential attachment (CPA) limit ($G(k) = k$ and $p(t) = b$) of our model dictates that the asymptotic scaling exponent should be $\gamma_{\mathrm{CPA}} = (2-b)/(1-b)$. Since the data will seldom have reached their asymptotic regime, deviations will be recorded and the connection between α and γ will be partly lost. To obtain asymptotic results for growth functions that are not strictly linear for all values of k, one must study each scenario on a case-by-case basis [67, 36]; estimating κ alone requires the integration of the model. Nevertheless, despite the absence of exact expressions for $p(t)$ and $G(k)$, the flexibility of the derived functional form will provide a useful and versatile parametrization of the complete temporal evolution of empirical data. The results of the next sub-sections confirm this assertion.

3.2.3 Reconstructing the past

The model based on Eq. (3.2) may now be used to replicate the growth of empirical distributions. Our objective is in part to verify the presence of constraints on the birth, Eq. (3.17), and growth, Eq. (3.5), of individuals; but also to use these constraints to determine the past and future of different systems solely from a snapshot of their present distribution.

Our model consists of iterating Eq. (3.2) for all k, with a given combination of $p(t)$ and $G(k)$, until time t reaches the total resource, K, of the system's present state. Hereafter, *we do not at any point show actual fits of the temporal data*, but instead find the optimal combination of $p(t)$ and $G(k)$ that minimizes the error produced by Eq. (3.2) when modelling the present state of a given system.

A simple analogy will clarify first the strategy behind our optimisation procedure. We are given a semi-infinite vertical chain of buckets. At the bottom of each one we drill a small hole of various width such that the k-th bucket has a hole of size $G(k)$. The first bucket, at the top of the chain, is placed under a dripping faucet whose flow is controlled in time by the function $p(t)$. Our goal is to adjust both the flow of the water $p(t)$ and the width of the holes $G(k)$ in order to reach a target quantity $\tilde{N}_k(t_f)$ of water for each bucket k after a time t_f. This target quantity is itself produced by a hidden $\tilde{p}(t)$ and $\tilde{G}(k)$. Since the function $G(k)$ has an infinite number of degrees of freedom, this means that for almost any $p(t)$ we could find a $G(k)$ respecting the target distribution. However, if the chosen $p(t)$ is very different from $\tilde{p}(t)$, the obtained $G(k)$ will also differ from $\tilde{G}(k)$. Therefore, we constrain $p(t)$ first, having a few degrees of freedom, before optimizing $G(k)$ accordingly.

The quality of our model representation $[p(t), G(k)]$ is assessed by counting the number of individuals $\{N_k(t_f)\}$ (or water drops) assigned to the wrong share k (or the wrong bucket) with respect to the empirical state $\{\tilde{N}_k(t_f)\}$,

$$\Delta\left[p(t), G(k)\right] = \frac{1}{2}\sum_k \left|\tilde{N}_k(t_f) - N_k(t_f)\right| . \tag{3.18}$$

A number of points are worth mentioning. Firstly, the measure Δ, based on absolute errors, was chosen over, say logarithmic or cumulative errors, because of its robustness to the tails

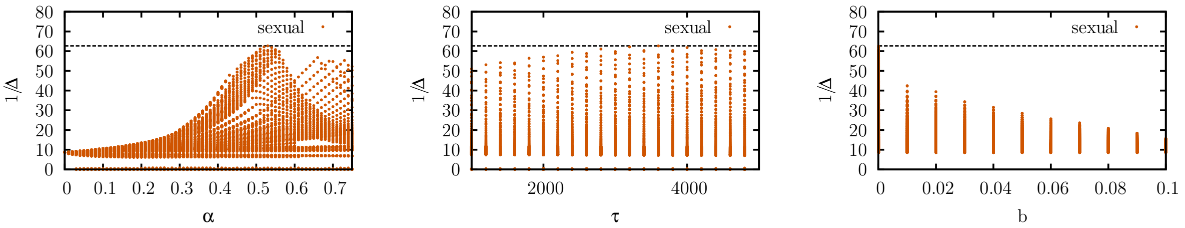


Figure 3.2 – **Parameter sweep.** Quality of our ability to model the growth of the database of sexual activities with $G(k) = k$ and various $p(t)$. The quality measure is given by $1/\Delta$ (see Eq. 3.18) and its maximal values are indicated with a dotted line at $1/\Delta = 62.68$ corresponding to 1.6% of misassigned shares $k_i(t_f)$ at $\alpha = 0.53$, $\tau = 3600$ and $b = 0$. Note that these figures are projections of a 3 dimensional fitness landscape.

of the distributions where the finite-size data falls to a non-zero value ($\propto N(t_f)^{-1}$) while the mean-field model falls to zero. Secondly, although minimisation of Δ (or optimisation of $[p(t), G(k)]$) is conducted on the sole knowledge of the present state of the system, i.e. $\{\tilde{N}_k(t_f)\}$, our model completely reconstructs its pre-history. Thirdly, while the search for the optimal parameter values of $p(t)$ seems a daunting enterprise, a number of internal and empirical restrictions on $p(t)$ constrains the quest: i. since $p(t) \in [0, 1]\ \forall\ t$, $b \in [0, 1]$ and therefore $-b \leq a(t + \tau)^{-\alpha} \leq (1 - b)$; ii. since $p(t) = \dot{N}(t)$ by definition, the total empirical population $\tilde{N}(t_f)$ can serve as normalisation, removing one degree of freedom:

$$a = \frac{\tilde{N}(t_f) - bt_f}{(t_f + \tau)^{1-\alpha} - (1 + \tau)^{1-\alpha}} (1 - \alpha) \ . \tag{3.19}$$

Because a can be positive or negative, our model can just as well describe a growing or decreasing birth function. Finally, the optimisation procedure is carried out in two stages: i. a close set of optimal triplets $[\alpha, \tau, b]$ are obtained by scanning parameter space to minimise Δ while maintaining initially $G(k) = k$ (Fig. 3.2 presents an example of this parameter scan); ii. the growth function $G(k)$ is allowed to vary under the newly acquired best possible $p(t)$ and constrained by the empirical data $\{\tilde{N}_k(t_f)\}$. Details of the algorithm are given in Sec. 3.4. Based on the quality of the obtained model $[p(t), G(k)]$, no further optimisation was found necessary.

While the systems studied in Fig. 3.3 vary in nature, age and distributions, our results indicate that they follow qualitatively the same evolution, and confirm the presence of both a delayed regime of temporal scaling and preferential attachment in all cases. Point estimates (Maximum-Likelihood Estimation (MLE) over the binary sequence of birth and growth events, see Sec. 3.4) of the relevant parameters are given on Table 3.1 and are visually compared with our model in Fig. 3.3(left). The behaviours extracted by our model from static distributions (without temporal data) are thus shown to be good estimates of the best possible fits to the actual temporal data.

Because of the form $p(t) = a(t+\tau)^{-\alpha}+b$, the complementary probability (i.e. the probability

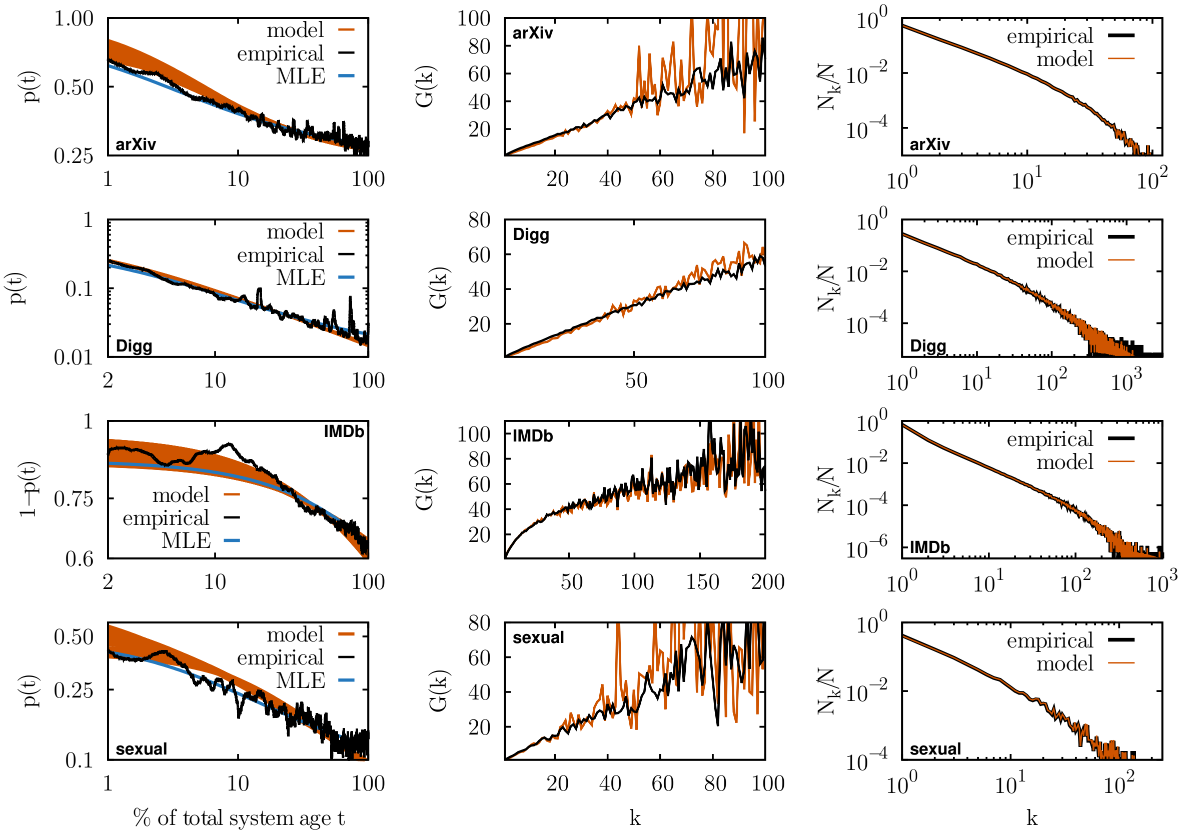


Figure 3.3 – **Temporal scaling and preferential attachment in human systems.** From left to right: birth function with temporal scaling of the form $a(t+\tau)^{-\alpha}+b$; growth function with asymptotic preferential attachment; scale-independent distributions. (left) The orange curves represent birth functions leading to predictions within 25% of the minimal error between model and empirical data (present state only). The empirical black curves are presented solely for comparison as no temporal data is needed for our reconstruction of the past. Likewise, Maximal-Likelihood Estimates (MLE) of $p(t)$, calculated *with* the actual sequence of birth and death events are shown in blue to highlight the accuracy of our model. (middle and right) Growth functions and present distributions: only the curves with the absolute minimum error are shown. The systems are, from top to bottom: distribution of papers per author in the arXiv [$N(t_f) = 386,267$ at $t_f = 1,206,570$], votes per user on Digg [$N(t_f) = 139,409$ at $t_f = 3,018,197$], movies per actor on IMDb [$N(t_f) = 1,707,525$ at $t_f = 6,288,201$] and relations per individual in the sexual data [$N(t_f) = 16,730$ at $t_f = 101,264$]. The methodology to measure the empirical birth and growth functions is presented in Sec. 3.4.

Table 3.1 – MLE point estimates of parameters using the empirical sequence of birth and growth events.

system	arXiv	Digg	IMDb	sexual
α	0.58	0.95	0.46	0.60
τ	$12,066$	$60,364$	$6,288,202$	$3,038$
b	0.240	0.012	0.976	0.072

that the t-th event is a growth event) has the same form with $a' = -a$ and $b' = 1-b$. This fact is highlighted with the case of IMDb in Fig. 3.3 and is consistent with our analysis where the constant a (but not b) can be negative. Furthermore, notice that IMDb is not only the sole system for which $p(t)$ is an increasing function, but also the only system for which $G(k)$ has initially a non-linear behaviour, and consequently a large τ. This confirms our interpretation of the role of τ as a coupling between population growth, $p(t)$, and individual growth, $G(k)$. With hindsight, this initial regime of the IMDb growth function probably corresponds to the so-called *star system*: actors with little experience are far less likely to be chosen for a role than experienced actors, but the first few movies in a new actor's curriculum are also far more important than the n-th in the career of a well-established star. This influences the introduction rate of new actors to preserve the system's scale independence. This interpretation is somewhat speculative, yet the fact remains that these effects are observed in the temporal data and that our model is able to extract them solely from the present distribution.

With the exception of one much smaller system (sexual data), the quality of our reconstruction of the past is surprisingly good considering that it requires no temporal data whatsoever. For instance, the Digg user activity distribution led us to determine with very high precision that 25% of votes are due to new users 12 hours into the month, whereas this proportion falls below 2% by the end of the month.

Our ability to infer the birth function based on a single snapshot also implies that we can distinguish between systems close or far from equilibrium (i.e. their statistical steady-state). For all investigated cases, both the inferred and observed $p(t)$ agree that none of these systems have reached their asymptotic b value. In the Digg database, it is even unclear if this value exists at all. In other systems, it is interesting to discern whether the distribution is approaching its asymptotic scale exponent γ from above (less heterogeneity) or below (more heterogeneity). For instance, the sexual database describes a network for which the first two moments of the activity distribution determine whether or not the introduction of a given sexually transmitted infection will result in an epidemic [48, 62]. These moments being defined by the scale exponent, our ability to describe the system's approach to equilibrium directly translates in an ability to determine which infection could invade the network.

More generally, this idea leads to a crucial point. The results confirm that our model embodies

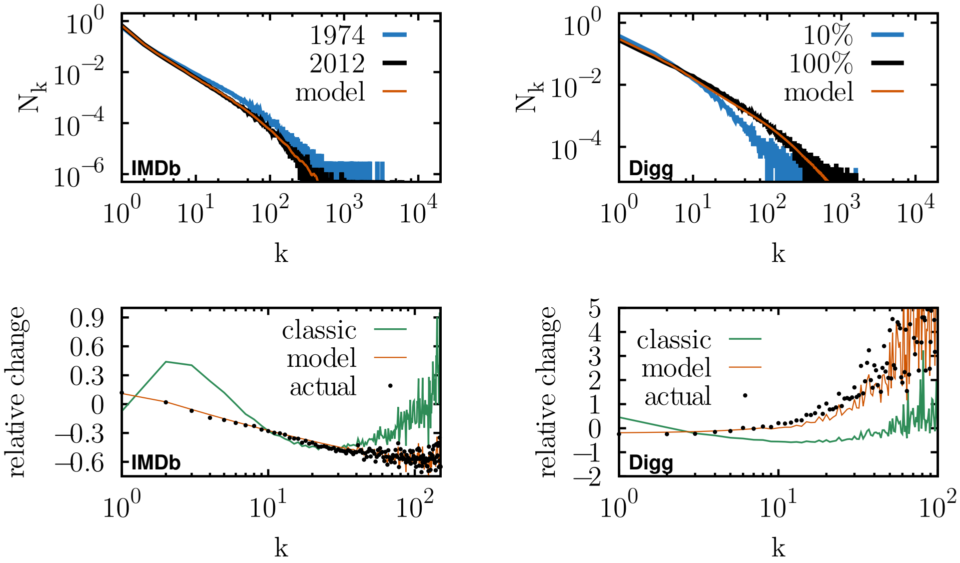


Figure 3.4 – **Prediction of present state from a snapshot of a past distribution.** The model uses only the distribution at $t_i = 0.3t_f$ (IMDb) and $t_i = 0.1t_f$ (Digg) of the system's history (in blue) to reconstruct the past (i.e. the birth and growth functions) and predict the future (in orange) of the database (in black). (top) Past, present (actual and predicted) distributions. (bottom) Relative change of each compartment N_k measured as $[N_k(t_f) - N_k(t_i)]/N_k(t_i)$; where $N_k(t_f)$ is either the actual distribution or a prediction. For comparison, a prediction using the classic preferential attachment model [92, 56] is shown in green.

intrinsic constraints of scale independence. These constraints appear to clearly define the possible paths that a system can follow. A snapshot of its present state is then sufficient to determine where it comes from and where it is heading. This naturally leads to a second question: can we use the *reconstructed* past of a system to *predict* its future?

3.2.4 Predicting the future

To turn our model into a predictive tool is a simple matter. We first eliminate the statistical fluctuations present in the reconstructed growth function. It is reasonable to assume that these fluctuations stem not from the form of the growth function itself but merely from the system's finite size and the stochastic nature of the dynamics. The fluctuations are eliminated by applying a linear fit to the asymptotic behaviour of the reconstructed $G(k)$. A prediction can then be obtained by iterating Eq. (3.2) from a chosen present state to a desired future time.

We apply this predictive model to the largest databases, i.e. actor productivity in the IMDb and user activities on Digg. The results are shown in Fig. 3.4(top). By using the activity distribution on Digg after only three days (again without any temporal data, only the current activity distribution per user), we can extrapolate the distribution over the period of a month.

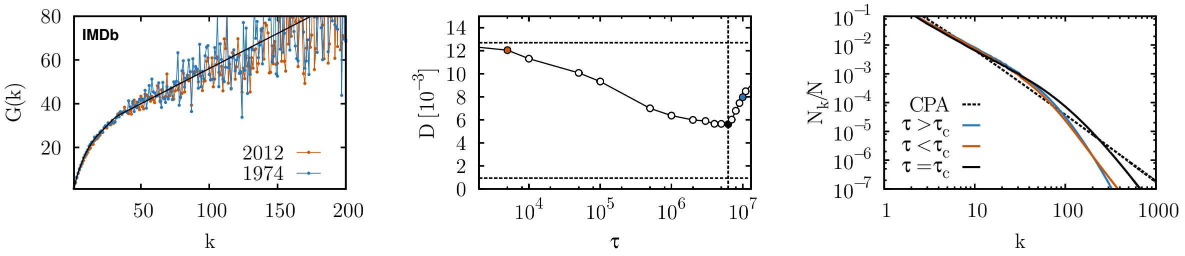


Figure 3.5 – **Robustness (and sensitivity) of the mechanism and model.** (left) The growth function inferred on the full IMDb dataset (orange), as shown in Fig. 3.3, is compared with the function inferred with 30% of IMDb's history (blue) as used in Fig. 3.4. The black curve is the smooth version used to predict IMDb's future. (middle) The smooth growth function of IMDb is used with different $p(t)$ to obtain distributions and measure their distance to a true power-law behaviour. The lower the distance, the closer the model is to scale independence. The upper horizontal dotted line corresponds to $p(t) = \langle p(t) \rangle$ with IMDb's smooth growth function. The lower horizontal dotted line corresponds to classical preferential attachment: $p(t) = \langle p(t) \rangle$ and $G(k) = k$. With IMDb's growth function, the minimum distance (the most power-like behaviour) is indicated with the vertical dotted line at 6.25×10^6 ($\pm 2.5 \times 10^5$) in close agreement with the MLE values of Table 3.1. (right) Examples of the distributions obtained with different values of τ are compared to the classical preferential attachment (CPA) which ignores the system's intrinsic $G(k)$ by using $G(k) = k$. The color code follows the color coded dots of the middle figure.

In contrast, assuming a constant birth rate (as in classical preferential attachment [92, 12, 56]) leads to a predicted final population of 475,000 users. Our model correctly compensates for repeated traffic and predicts a population of 115,000 users, closer to the correct value of 139,000 and missing only some sudden bursts of new user influx. This observation embodies the strength of our model and the importance of a time dependent birth rate. Similar results are obtained for actor productivity on the IMDb. Remarkably, we reproduce the state of the system at year 2012 from its state at year 1974. Given that extrapolation is a delicate procedure, it seems not unlikely that these agreements are not coincidental. As a comparison, the classical preferential attachment model shown in Fig. 3.4(bottom) is incapable of discerning whether the scaling exponent of a system is increasing or decreasing with time. Since the classic model ignores the temporal dependency introduced here, our results highlight the importance of linking the temporal and organizational features of complex systems.

It could be argued that the growth function should more generally depend on time to include potential changes in mechanisms. However, our ability to predict the future with a time-independent growth function seems to rule out, at least in the cases studied, the necessity for a temporal dependence. In fact, Fig. 3.5(left) compares the growth function inferred from the IMDb using only records before 1974 and before 2012. While the dataset has more than tripled in size during these 40 years, the inferred growth functions do not significantly differ from one another, thereby explaining the quality of our results shown in Fig. 3.4. This also implies that although the growth function has an influence on the time dependency of the dynamics (through the coupling parameter, or delay, τ), it does not itself depend on time. This is particularly surprising considering that the movie industry has changed dramatically between these two snapshots. One recalls that 1975 saw the rise of the Blockbuster era following the release of Steven Spielberg's *Jaws* [75]. The following change in movie making did not affect the dynamics of the system, which suggests that the growth function may be intrinsic to the considered human activity and robust to environmental or societal changes. The growth function of Digg is similarly robust through time as the dataset spans a single month of activity. While generalizations of our model could be considered, with growth functions varying in time or across individuals [19], the apparent time independence of the growth function is surely worthy of future investigations. Contrariwise, were the mechanism(s) of a system growth function to change over time, this would reflect immediately in our inability to predict the future and would precisely be an indication of changes in the underlying mechanism(s). Hence, even if it was to fail, this model would offer significant insights.

3.2.5 Coupling between the growth function and the temporal delay

An important insight of the previous analysis states that the delay τ embodies an inherent coupling between the growth function $G(k)$ and the birth function $p(t)$ to ensure robust scale independence. Put differently, any non-linearity of $G(k)$ for small k should be compensated by the temporal delay τ if the system is to be roughly scale-independent even for small time t.

In order to examine this assertion, we make the following experiment. We use IMDb's growth function as it is highly non-linear for small k, and test the plausibility of a power law fit to the model for different $p(t)$. We fix the temporal scaling α to IMDb's 0.55 0.55, and we fix the value of a and b by setting both $p(1)$ and the average $\langle p(t)\rangle$ (for $t \in [1, 5 \times 10^6]$) also to that of IMDb. The only parameter allowed to vary freely is the temporal delay τ. Hence, we always have the same population growing with the same growth function for the same number of time steps, and starting with the same initial birth rate but with different delays τ between the initial and final regime of $p(t)$.

We then iterate Eq. 3.2 with each $p(t)$ to obtain the distribution N_k/N from which we randomly generate ten populations of size $N(t)$ to emulate a real system of finite size. The generated data is then fitted to a power-law distribution with the method of Clauset, Shalizi and Newman [28]. The quality of the power-law hypothesis is finally measured with the distance between the fitted power-law distribution N_k^*/N and the original distribution N_k/N obtained from the model. This distance D is calculated through the Jensen-Shannon divergence of the two distributions and averaged over the ten generated populations, see Sec. 3.4 for details. This approach provides an estimate of how surprising it would be for a sample obtained from our distributions to have been produced by an actual power-law distribution.

The results highlight that, given IMDb's growth function, the particular $p(t)$ which was observed in the temporal data of IMDb and obtained from our algorithm is the most robust way for this system to grow towards scale independence. In other words, the $p(t)$ observed in the IMDb effectively compensates the non-linear deviation observed in its growth function in a way that ensures a fast convergence to scale independence. Figure 3.5(right) illustrates this, comparing three distributions obtained with different $p(t)$ with the classical preferential attachment ($[p(t) =< p(t) >, G(k) = k]$). The distribution obtained with the optimal solution ($\tau = \tau_c$) is clearly already ahead of the other, and not so far from the CPA, on the path to scale independence.

In a nutshell, this simple experiment adds further strength to the validity of our theoretical framework, and reasserts one of its important conclusions: arbitrary growth rules do not all lead to scale independence, and certainly not all at the same speed. Finally, while we confirmed our theoretical insights and our ability to use them in practical applications, the mechanisms by which $p(t)$ might self-organize in these systems to assure scale independence remain unknown.

3.3 Discussion

In this study, instead of directly studying the classical preferential attachment model, we have derived a more general form from the simple assumption that a power-law distribution is a good *approximation* of a distribution of interest. Our general model differs from the classic idealized version in two ways: the growth (or attachment) function is given some flexibility in its initial behaviour, only required to be asymptotically linear; and the birth function is *time dependent* through a delayed temporal scaling. This delay acts as a coupling between two levels of dynamics: the growth of the population and the growth of a given individual's activity.

This general model is both flexible and constrained enough to be useful. In fact, we have shown that a three dimensional parameter space (temporal scale exponent, delay and asymptotic birth rate) is sufficient to infer the past and future of a present distribution.

It is important to keep in mind that our analysis is in no way restricted by the nature of the systems under study. Considering that scale-independent systems are ubiquitous in science and everyday life, but that temporal data on their growth is seldom available, our framework provides a new investigation line to reconstruct their past and to forecast their future.

3.4 Appendix 3.A: Methods

3.4.1 Description of databases

Prose samples. Text files for the works of William Shakespeare, Miguel de Cervantes Saavedra and Johann Wolfgang von Goethe were downloaded from the Project Gutenberg at `www.gutenberg.org/`. Punctuation marks and Project Gutenberg disclaimers were removed from the files manually.

While not a human system, but certainly a man-made one, these prose samples were used to get better statistics on the birth function. While human systems are unique and time dependent, written texts feature a translational invariance [17]. This property allows us to gain better statistics of their growth by considering multiple samples of equal length as different realizations of the same process.

Time t and resource $K(t)$ correspond to the total number of written words. Individuals correspond to unique words and their share $k_i(t)$ to their number of occurrences.

Scientific authorships on the arXiv. This database consists of a chronological list of all author names appearing on papers of the arXiv preprint archive (in order of publication date). It was compiled using the arXiv API to gain a full list of scientific publications available from `http://arxiv.org/` as of April 2012.

Time t and resource $K(t)$ correspond to the total number of paper authorships. Individuals correspond to authors and their share $k_i(t)$ to their number of publications.

Digg user activities Digg (`http://digg.com/`) is a social news website where registered users can vote on news or other types of articles that they deem interesting. This database is a list of all user votes on top stories (frontpage) over a period of one month in 2009 [70].

Time t and resource $K(t)$ correspond to the total number of votes. Individuals correspond to registered users and their share $k_i(t)$ is their respective number of votes.

IMDb castings The Internet Movie Database (`http://www.imdb.com/`) consists of an impressive amount of cross referenced lists (released films, cast and crew, etc.). These databases can be accessed or downloaded in various ways: see `http://www.imdb.com/interfaces` for details. From the list of actors featured on IMDb, which records all movies in which they have appeared, and the list of movie release dates, we built the chronological sequence of 'castings'.

Time t and resource $K(t)$ correspond to the total number of castings (a given actor playing in a given film). Individuals correspond to unique actors and their share $k_i(t)$ is the total number of films in which they have appeared.

Sexual activities in a Brazilian community This database was built from a public online forum for male clients who evaluate relations with female prostitutes [30]. After preliminary results using the client and prostitute databases separately, we concluded that it was not necessary to distinguish between the two. The simplified database is thus a list of unique IDs corresponding to either a client or a prostitute, in chronological order of sexual relations (at time of online posting).

Time t and resource $K(t)$ correspond to the total number of such IDs (twice the total number of relations). Individuals correspond to unique IDs (either client or prostitute) and their share $k_i(t)$ is their respective number of relations.

Table 3.2 – Summary of database sizes and quantities.

Quantities	prose	arXiv	Digg	IMDb	Sexual
Individuals	unique words	authors	users	actors	clients/prostitutes
$N(t_f)$	502 on average	386,267	139,409	1,707,565	16,730
Resource	written words	papers	votes	castings	sexual activities
$K(t_f) = t_f$	cut at 1000	1,206,570	3,018,197	6,288,201	101,264

3.4.2 Measuring the birth function

Prose samples The translational (or temporal) invariance of written text implies that we can consider different samples of equal length from the same author as different realizations of the same experiment. The files were thus broken into samples of equal length and analysed separately. Each experiment can be reduced to a binary sequence of ones (when the word is a new word; i.e. a birth event) and zeros (when the word is an old one; a growth event). The birth function $p(t)$ of a given author can then be obtained by simply averaging all binary sequences.

Other systems In the other systems, since preliminary tests excluded the possibility of temporal invariance, a different procedure was used. The simplest one is to merely apply a running average on the binary sequence of birth and growth events. We used temporal windows of Δt equal to 1% of the total system size (final time t_f) for the two largest databases (Digg and IMDb) and between 0.5% and 1% of system size for the others. This method was shown to preserve the delayed temporal scaling on a random binary sequence whose elements were drawn from a known probability distribution following $p(t)$.

3.4.3 Measuring the growth function

In this section, we detail the procedure used to obtain the growth function $G(k)$ of a system from its temporal data, $t \in [0, t_f]$. We use the following notation: we keep in memory every encountered individual i, its number of appearances (or current share) $k_i(t)$, $N_k(t)$ as the number of individuals with share $k_i(t) = k$ and the total population $N(t)$ after time t. Starting from $t = 1$, we proceed as follows.

1. *Event.* If the t-th individual is new, add it to memory and note:

$$\begin{aligned} N(t) &= N(t-1)+1 \\ k_{N(t)}(t) &= 1 \\ N_1(t) &= N_1(t-1)+1 \end{aligned}$$

and go to step 4. If the individual is already in memory, go to step 2.

2. *Chances.* We increment a function of chances

$$C(k,t) = C(k,t-1) + N_k(t-1)/N(t-1) \quad \forall\, k$$

and go to step 3.

3. *Success.* Noting i the individual involved in the current event, increment a function of successes

$$\begin{aligned} S(k_i(t-1),t) &= S(k_i(t-1),t-1)+1 \\ S(k,t) &= S(k,t-1) \qquad \forall\, k \neq k_i(t-1) \end{aligned}$$

and the following variables

$$\begin{aligned} k_i(t) &= k_i(t-1)+1 \\ N_{k_i(t-1)}(t) &= N_{k_i(t-1)}(t-1)-1 \\ N_{k_i(t)}(t) &= N_{k_i(t)}(t-1)+1 \end{aligned}$$

and go to step 4.

4. *Repeat.* If we have reached the end of the database, go to step 5. Otherwise, go to step 1.
5. *Calculation.* The growth function is finally given by:

$$G(k) = S(k,t_f)/C(k,t_f) \quad \forall\, k$$

corresponding to the ratio of actual successes to chances under a uniform growth. The deviation from $G(k) = 1\ \forall k$ is the actual growth function.

3.4.4 Reconstructing the empirical growth function

Once the best possible $p(t)$ has been found, we adjust the growth function $G(k)$ by iterating the following algorithm:

1. *Initialization.* We fix $p(t)$ and we first consider $G(k) = k$.
2. *Growth.* We iterate the following equation from $t = 1$ with $N_k(1) = \delta_{k1}$ up to t_f:

$$N_k(t+1) = N_k(t) + p(t)\delta_{k1} + \frac{1-p(t)}{\sum G(k)N_k(t)}\left[G\left(k-1\right)N_{k-1}(t) - G(k)N_k(t)\right] \ .$$

3. *Correction.* For all k, we adjust $G(k)$:

$$\overline{G}(k) = G(k)\frac{N_k(t_f)/\sum_{i=k}^{\infty} N_i(t_f)}{\tilde{N}_k(t_f)/\sum_{i=k}^{\infty} \tilde{N}_i(t_f)}$$

4. *Iteration.* We set $G(k) = \overline{G}(k)$ and return to step 2.

At step 3, the adjustment factor is simply the ratio of "the quantity of individuals (water) that made it to share (bucket) k but did not go to $k+1$", as calculated in the model $N_k(t_f)$ versus the target distribution $\tilde{N}_k(t_f)$. This algorithm is usually iterated 4 or 5 times to obtain a converged growth function.

3.4.5 Maximum-likelihood estimation

We search for a $p(t)$ that maximizes the binary logarithm of the likelihood $\mathcal{L}$ of a given binary sequence $\{y_i\}$ of birth ($y_i = 1$) and growth events ($y_i = 0$):

$$\log_2 \mathcal{L}\left(\tau, \alpha, b \mid \{y\}\right) = \sum_{i=1}^{t_f} y_i \log_2 p(i) + (1 - y_i) \log_2 \left(1 - p(i)\right) \ .$$

3.4.6 Jensen-Shannon divergence

Given two distributions, $\boldsymbol{M}$ and $\boldsymbol{F}$, with probabilities $\{M_i\}$ and $\{F_i\}$ respectively, the quantity

$$D_{\text{KL}}\left(\boldsymbol{M} \| \boldsymbol{F}\right) = \sum_i M_i \log_2 \left(\frac{M_i}{F_i}\right) \tag{3.20}$$

is called the *Kullback-Leibler distance* [88] between $\boldsymbol{M}$ and $\boldsymbol{F}$, or the relative entropy between the two distributions. A close relative of this quantity, also referred to as the *Jensen-Shannon divergence*, is a symmetric form given by

$$D_{\text{SKL}} = \frac{1}{2} D_{\text{KL}}\left(\boldsymbol{M} \| \boldsymbol{A}\right) + \frac{1}{2} D_{\text{KL}}\left(\boldsymbol{F} \| \boldsymbol{A}\right) \tag{3.21}$$

where the distribution $\boldsymbol{A}$ with probabilities $A_i = \left(M_i + F_i\right)/2$ is used to approximate $\boldsymbol{M}$ or $\boldsymbol{F}$ respectively.

In our study, we want to quantify the similarity between the distribution, $\boldsymbol{M}$, generated by our mean-field model and the distribution $\boldsymbol{F}$ obtained from a corresponding power-law fit. In practice, the procedure goes as follows: with the distribution $\boldsymbol{M} = \{N_k/N\}$, we generate a number of population samples $\{m^{(j)}\}$ of size $N(t_f)$ and fit each of them to a power-law $f^{(j)}$ using the standard method of Clauset et al. [28]. Each $f^{(j)}$ is characterized by an exponent $\gamma^{(j)}$ and a minimal value $k_{\text{min}}^{(j)}$ (here always equal to 2) marking the beginning of the power-law tail. These power-law populations are then used to construct the related distributions $\left[\boldsymbol{F}^{(j)} = \{N_k^{(j)}/N\}\right]$ which are finally compared to the tail of the original distribution $\boldsymbol{M}$ over the range $k_{\text{min}}^{(j)} \leq k \leq 5000$ [$\sim$ IMDb's $k_{\text{max}}(t)$]. The comparison is quantified through the symmetrical Kullback-Leibler distance averaged over the different samples

$$D\left(\boldsymbol{M}, \boldsymbol{F}\right) = \langle D_{\text{SKL}}\left(\boldsymbol{M}, \boldsymbol{F}^{(j)}\right)\rangle_j . \tag{3.22}$$

3.5 Appendix 3.B: Internal validation of the framework

This section investigates in detail the behaviour of our general preferential attachment process. More precisely, we validate analytical results on both the time evolution of the system and on its asymptotic state.

3.5.1 Tracking a general preferential attachment process

We here wish to validate some of the secondary results derived form our framework, namely:

$$\sum_k G(k) N_k(t) = \kappa(t+\tau) \; , \tag{3.23}$$

$$k_{\text{max}}(t) \propto (t+\tau)^{(1-b)/\kappa} \; , \tag{3.24}$$

both in the limit $t \gg 1$ and $k_{\text{max}}(t)$ greater than some bound k^* to be identified.

To this end, we will consider three growth functions $G(k)$. First, the classical preferential attachment case; second, a concave function similar to the one observed in IMDb; and third, a convex functions describing an (as yet unobserved) increasing returns behaviour for small k. Mathematically,

$$G_{\text{CPA}}(k) = k \; , \tag{3.25}$$

$$G_{\text{conc}}(k) = k + \sqrt{k} \; , \tag{3.26}$$

$$G_{\text{conv}}(k) = \left(\frac{k}{10}\right)^2 e^{-k/10} + \left(1 - e^{-k/10}\right) k \; . \tag{3.27}$$

The convex function quickly (exponentially) converges towards a linear behaviour, whereas we have chosen a much slower convergence ($1/\sqrt{k}$) for the concave function. As the concave case was previously studied within the IMDb dataset, we will here try to investigate how far k^* might be for slower convergence to linearity. Also, note that for all growth functions we have $G(k) = k$ in the limit $k \to \infty$, as requested in our derivation of $k_{\text{max}}(t)$ (to simplify the calculation of κ). Otherwise, if $G(k) = a_0 k$ for instance, the $1/\kappa$ in Eq. (3.24) would actually be a_0/κ. Since $G(k)$ can be rescaled arbitrarily, we fix its linear slope to unity. The three growth functions are illustrated in Fig. 3.6.

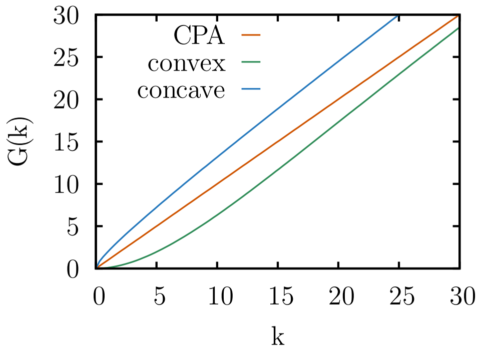


Figure 3.6 – **Three investigated growth functions.** Plots of the growth functions $G_{\mathrm{CPA}}(k)$, $G_{\mathrm{conc}}(k)$ and $G_{\mathrm{conv}}(k)$ described in the text.

Normalisation of growth probabilities

Considering a constant birth rate, $p(t) = 1/3$, we follow $\sum_k G(k)N_k(t)$ as the general PA model is iterated with different $G(k)$. The results are shown on Fig. 3.7. The results are not surprising considering the form of the sum. Considering the linear behaviour (with slope normalized to unity) of $G(k)$ starting at some appropriate bound k^*, we write:

$$\begin{aligned}\sum_i G\left(k_i(t)\right) = \sum_{k=1}^{k_{\max}(t)} G(k)N_k(t) &= \sum_{k=1}^{k^*-1} G(k)N_k(t) + \sum_{k=k^*}^{k_{\max}(t)} G(k)N_k(t) \\ &= \sum_{k=1}^{k^*-1} G(k)N_k(t) + \sum_{k=k^*}^{k_{\max}(t)} kN_k(t)\ . \end{aligned} \tag{3.28}$$

Using our definition of time,

$$\sum_{k=1}^{k_{\max}(t)} kN_k(t) = t \quad \rightarrow \sum_{k=k^*}^{k_{\max}(t)} kN_k(t) = t - \sum_{k=1}^{k^*-1} kN_k(t) \tag{3.29}$$

such that

$$\sum_i G\left(k_i(t)\right) = t + \left\{\sum_{k=1}^{k^*-1} \left[G(k) - k\right] N_k(t)\right\}\ . \tag{3.30}$$

Now assuming that enough time has elapsed for $N_k(t) \simeq n_k N(t)$ to be a good approximation ($\{n_k\}$ begin the steady-state distribution). The sum becomes

$$\sum_i G\left(k_i(t)\right) \simeq t + N(t)\left\{\sum_{k=1}^{k^*-1} \left[G(k) - k\right] n_k\right\} \tag{3.31}$$

where the term in braces is time independent. As we can see, the leading temporal term in the sum is linear as the growth of $N(t)$ is *at most* linear (growing with a power $1-\alpha$ and converging to bt) in the limit of large time. Thus, the term proportional to $N(t)$ can act as an offset in the limit of large time (if non-linear), as a modifier of the slope of the sum (if linear)

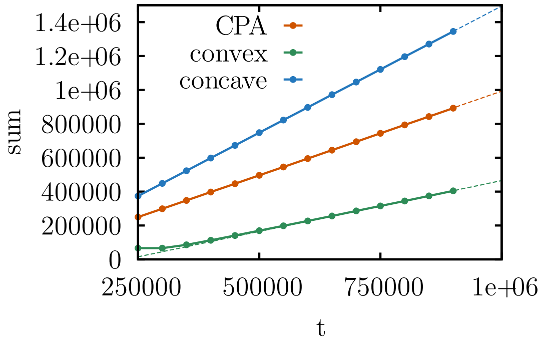


Figure 3.7 – **Normalisation of growth probabilities.** From top to bottom, the curves follow the sum $\sum G\left(k_i(t)\right)$ using the concave $G_{\text{conc}}(k)$, the classical $G_{\text{CPA}}(k)$ and the convex $G_{\text{conv}}(k)$. The linear fits are obtained using an implementation of the non-linear least-squares Marquardt-Levenberg algorithm.

or both. The main point is that its effect is also modulated by the sum $\sum_{k=1}^{k^*-1}\left[G(k)-k\right]$ which embodies the deviation of $G(k)$ from its linear behaviour in values $k<k^*$.

Figure 3.7 confirms this analysis. A $G(k)$ strictly equal to k will have slope strictly equal to unity: all events, birth or growth, have exactly a weight of one in the sum. A $G(k)$ initially below the line $G(k)=k$, like our convex function, will have a negative deviation from the linear behaviour: resulting in a slope smaller than unity, and in a negative offset. On the other hand, the opposite behaviour of our concave function implies a slope greater than unity and a positive offset.

Following the leader

The previous results confirm that we can use the form $\sum G(k)N_k(t)=\kappa(t+\tau)$ at least in some limit of large time $t \gg t^*$ (with t^* depending on the form of $G(k)$ and $p(t)$). Hence, our solution for the share k_{max} of the leading individual should also hold once $G(k_{\text{max}}) \sim k_{\text{max}}$ which happens above some bound $k_{\text{max}} > k^*$. Our solution,

$$k_{\text{max}}(t) = C_1(t+\tau)^{(1-b)/\kappa} \; , \tag{3.32}$$

is validated on Fig. 3.8. To track $k_{\text{max}}(t)$ in the iteration of our model, we simply scan values of k up to a fixed k_c (computational constraint); $k_{\text{max}}(t)$ is the value for which $\sum_{k=k_{\text{max}}(t)}^{k_c} N_k(t)$ is the closest to unity.

The analytical expressions on Fig. 3.8 (shown in dotted lines) use the values of τ and κ obtained from the fit on Fig. 3.7. For the concave, classical and convex growth functions, we found $[\kappa,\tau]=[1.4949,1]$, $[1,0]$ and $[0.66,-135000]$ respectively. The τ used to fit the $k_{\text{max}}(t)$ obtained with $G_{\text{conv}}(k)$ is modified to facilitate the fit. While the high value τ in the convex case is not surprising considering the highly non-linear initial behaviour of the growth

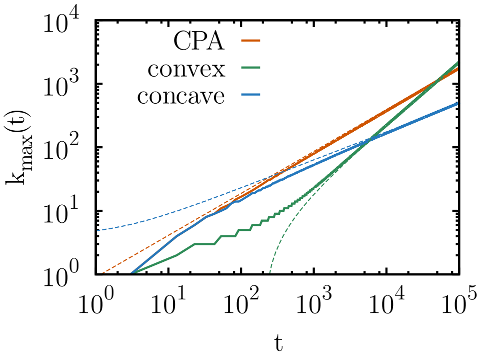


Figure 3.8 – **Share of the leading individual with different growth functions.** The different curves follow the same color code as before and use the same concave $G_{\text{conc}}(k)$, classical $G_{\text{CPA}}(k)$ and convex $G_{\text{conv}}(k)$. The solutions (dotted lines) follow $k_{\text{max}}(t) \sim (t+\tau)^{\beta}$ with $\beta = (1-b)/\kappa$ where $b = 0.33$ was fixed in our experiment and κ was obtained by fitting the results of Fig. 3.7.

function, it also means that tracking the $k_{\text{max}}(t)$ up to $t \gg |\tau|$ requires high k_c and is thus computationally impractical. Since, the scaling in $(1-b)/\kappa$ appears earlier, we simply divide τ by ten to compare qualitatively the two scaling regime. Similarly, the small value of τ obtained from the concave function is not surprising considering its behaviour is very close to $G(k) = k$, meaning the sum quickly converges to a linear function in Fig. 3.7. However, the square root correction in $G_{\text{conc}}(k)$ also means that one has to wait for a higher k^* until the ODE used to obtain $k_{\text{max}}(t)$ is valid. Thus, the weak non-linear behaviour leads to a small t^* in the convergence of the sum, but the slow convergence to an exactly linear behaviour in $G_{\text{conc}}(k)$ (as a square root versus as an exponential in $G_{\text{conv}}(k)$) implies a large k^*.

Scale independence of different snapshots

We apply the algorithm of Clauset *et al.* to infer the scaling exponents produced by the different growth functions at $t = 500,000$. Figure 3.9 presents the results of the algorithm. The only case where the algorithm clearly fails is the model grown with the convex growth function.

To illuminate the convex case, we push the iteration of the model further up to $t = 10^6$. The new results are shown on Fig. 3.10. The algorithm now performs much better, and we can also see why it initially failed. At $t = 10^6$, the expected population size is $N(t) = 0.33 \cdot 10^6 = 3.3 \cdot 10^5$. However, the highly non-linear behaviour of G_{conv} for small k forces a $k_{\text{min}} \sim 10^2$ for the obtained distribution. A visual inspection indicates that $\gamma_{\text{conv}} \sim 2$, such that we can only expect around 1% of the data points to fall after $k_{\text{min}} \sim 10^2$. This explains why the algorithm focuses on the initial behaviour of the distribution. With a bigger sample (i.e. a bigger population $N(t)$), there is little doubt that the algorithm would see the power-law tail. That being said, in all cases the algorithm has now converged to apparently better values.

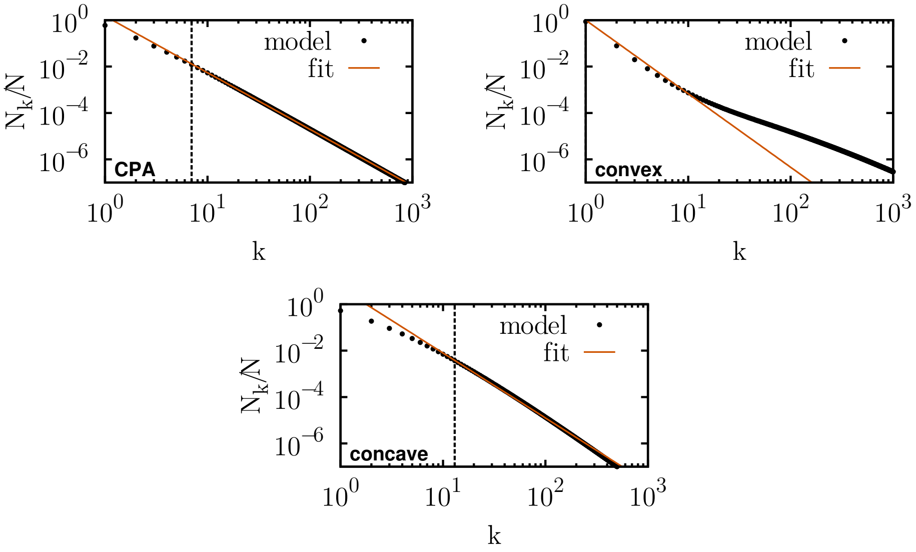


Figure 3.9 – **Algorithm of Clauset *et al.* applied to snapshot obtained with different $G(k)$.** For snapshots taken at $t = 500,000$, the inferred scale exponents are: $\gamma_{\text{CPA}} = 2.44$, $\gamma_{\text{conv}} = 3.17$ and $\gamma_{\text{conc}} = 2.80$. The theoretical steady-state CPA exponent is $(2-b)/(1-b) = 5/2$. The inferred k_{min} for which the power-law distributions *supposedly* holds are shown by the vertical dotted lines ($k_{\text{min}} = 1$ in the convex case).

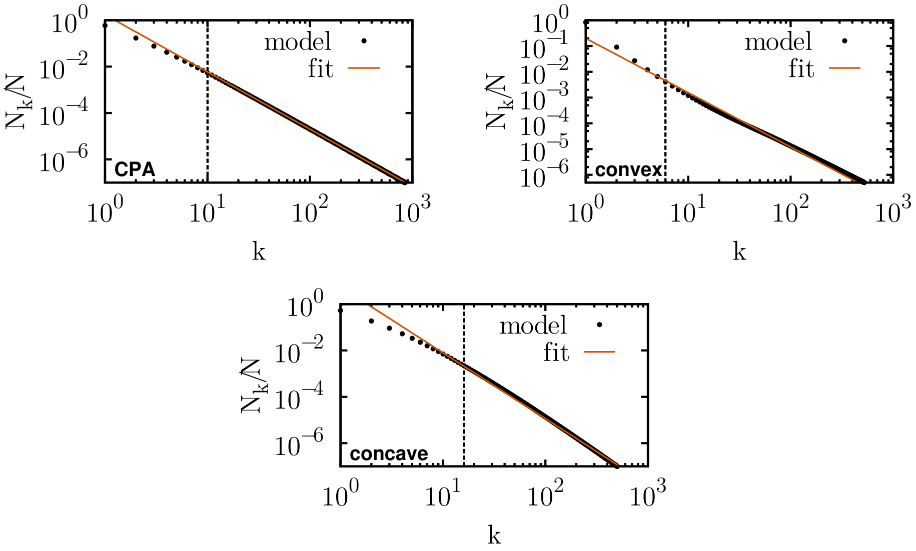


Figure 3.10 – **Algorithm of Clauset *et al.* applied to other snapshots.** For snapshots taken at $t = 1,000,000$, the inferred scale exponents are: $\gamma_{\text{CPA}} = 2.47$, $\gamma_{\text{conv}} = 2.11$ and $\gamma_{\text{conc}} = 2.84$.

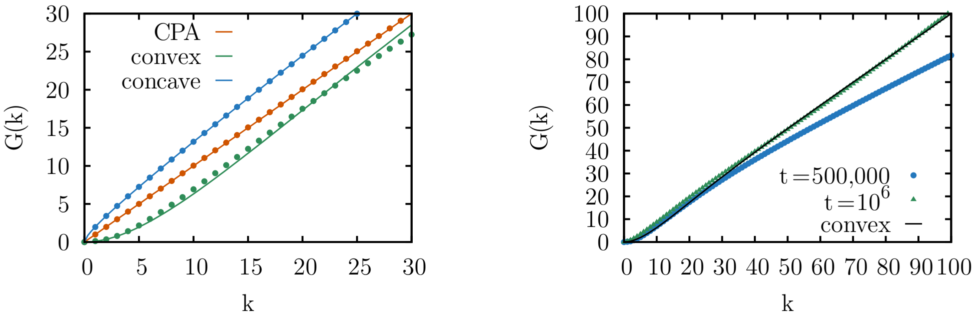


Figure 3.11 – **Reconstruction of synthetic growth functions with known past.** (left) The distributions presented in Fig. 3.9 are inserted in our algorithm for growth function reconstruction (along with the known $p(t) = 0.33$). The synthetic growth functions are represented with curves and the results of the algorithm with dots. After only 4 iterations (less for some), only the convex growth function is not perfectly recovered. (right) We compare the convex growth function with the one previously inferred after 4 iterations of our method on the distribution obtained at $t = 5 \cdot 10^5$ and one inferred after 10 iterations on the distribution at $t = 10^6$.

3.5.2 Testing the limits of our reconstruction algorithms

In light of the results obtained in the last subsection, particularly the peculiar form obtained with $G_{\text{conv}}(k)$ which is very convex for $k < 10$, we want to test our ability to reproduce the past of systems created by these growth functions.

Reconstructing the growth function

We now use the last set of distributions, obtained after $t = 5 \cdot 10^5$ time steps, and test whether or not our reconstruction of the growth function can be applied in all cases. Note that we first apply the optimization of $G(k)$ with a know $p(t)$. Figure 3.11(left) presents the growth functions inferred after a few iterations of the algorithm described in the main text. The algorithm performs well, except on the convex $G_{\text{conv}}(k)$.

To verify the robustness of our algorithm to convex growth function, we both consider a bigger system (now to $t = 10^6$) and do 20 iterations of the growth function reconstruction method. As observed on Fig. 3.11(right), the method is now more effective even though the convex regime is still not perfectly captured. The problem is then that that we quantify convergence of our method by the difference between successive iterations, but some growth functions cause a much slower convergence. It thus appears that a more conservative approach, i.e. systematically requiring a very low difference between iterations and/or a very high number of iterations, would be more appropriate when blindly applying the method on database with a completely hidden past.

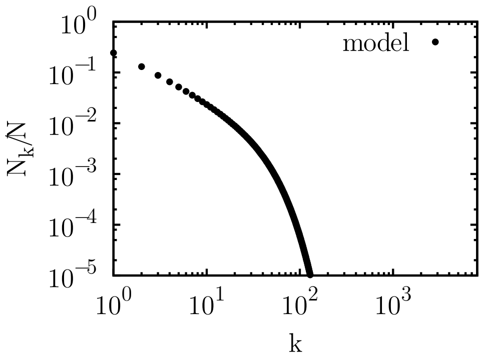


Figure 3.12 – **Reconstructing a far from scale-independent growth.** We produced a system with the concave growth function and $p(t) \propto (t + 10^4)^{-2/3}$. With a delay τ larger than necessary to compensate the non-linearity of the growth function, the produced distribution is very slowly converging to its scale-independent state.

Reconstructing an unknown and sub-optimal past

We model a system with a sub-optimal growth toward scale independence by using the concave growth function with a $p(t)$ using a τ much larger than the one estimated on Fig. 3.7. We use $p(t) = (t + 10^4)^{-2/3}(1 + 10^4)^{2/3}$ where the last factor is there to force $p(1) = 1$. After $t = 10^6$ time steps, we obtain the distribution shown in Fig. 3.12. Clearly, this $p(t)$ is sub-optimal and the system now converges much slower toward scale independence than observed in Fig. 3.9 and 3.10 which used a constant $p(t)$. Nonetheless, we will still attempt to reconstruct the past of this function based on this current snapshot.

As described before, we reconstruct the past by using $G(k) = k$ as a first approximation of the growth function and scanning the three dimensional parameter space for $p(t)$. We then quantify the quality of a given $p(t)$ by the sum of absolute errors between the target distribution and the tested model. *A priori*, we have no reason to expect the targets $[\alpha_T, \tau_T, b_T] = [2/3, 10^4, 0]$ (with which the target was produced) to yield the minimal error; but we at least want similar values to be in the list of potential candidates for the *most likely past*.

As before, we consider all $p(t)$ producing an error within a certain percentage of the absolute minimal error (tolerance) as candidates for the most likely past. The obtained estimates of $[\alpha, \tau, b]$ are given in Table 3.3. As expected, the minimal error is not obtained with the $p(t)$ used to produced the target as we are now using $G(k) = k$ as an approximation. The absolute error is found with $\alpha = \alpha_T$, $b = b_T$, but $\tau < \tau_T$. Yet, the correct $[\alpha_T, \tau_T, b_T]$ are within the intervals obtained for tolerance as low as 4%. Moreover, within 5% tolerance, we find two distinct local families of solutions: with $\tau \leq \tau_T$ and the correct $\alpha = \alpha_T$ or with $\tau > \tau_T$ and a $\alpha > \alpha_T$ to compensate.

Finally, as none of these solutions perfectly reproduce the target distribution, they can all be used to reconstruct the growth function starting from $G(k) = k$ and (rapidly) converging to $G_{\text{conc}}(k)$ as observed in Fig. 3.11. Unfortunately, reconstruction based on the convex $G_{\text{conv}}(k)$ were not investigated since a systematic way to deal with $\tau < 0$ has yet to be formulated. Most likely, one would need to start the model at $t > |\tau|$, but it is unclear what kind of initial conditions should be considered. Since, in this unique context, $k_{\text{max}}(t)$ as an initially slow growth, perhaps one could simply use a population of t individuals with share $k_i = 1$.

Table 3.3 – Candidates for the most likely past of the target distribution .

tolerance	α	τ	b
25%	$[0.66, 0.75]$	$[2000, 23000]$	$[0, 0]$
20%	$[0.66, 0.75]$	$[2000, 21000]$	$[0, 0]$
15%	$[0.66, 0.75]$	$[2000, 19000]$	$[0, 0]$
10%	$[0.66, 0.75]$	$[2000, 17000]$	$[0, 0]$
5%	$[0.66, 0.75]$	$[2000, 15000]$	$[0, 0]$
3%	$[0.66, 0.67]$	$[2000, 3500]$	$[0, 0]$

3.5.3 Revisiting the asymptotic state

Our model is essentially summarized by the following rate equation:

$$N_k(t+1) = N_k(t) + p(t)\delta_{k,1} + [1-p(t)]\,\frac{N_{k-1}(t)G(k-1) - N_k(t)G(k)}{\sum N_{k'}(t)G(k')}\ . \tag{3.33}$$

Using our functional form for $p(t)$ and a continuous time approximation yields

$$\frac{d}{dt}N_k(t) = \left[a(t+\tau)^{-\alpha} + b\right]\delta_{k,1} + \left[1 - a(t+\tau)^{-\alpha} - b\right]\frac{N_{k-1}(t)G(k-1) - N_k(t)G(k)}{\sum N_{k'}(t)G(k')}\ . \tag{3.34}$$

Using a continuous approximation in k (i.e. $N_k(t) \to N(k,t)$) as we did in Sec. 2.2.3, we can rewrite the last equation as the following system:

$$\frac{\partial}{\partial t}N(1,t) = \left[a(t+\tau)^{-\alpha} + b\right] - \left[1 - a(t+\tau)^{-\alpha} - b\right]\frac{N(1,t)G(1)}{\sum N(k',t)G(k')} \quad \text{for } k = 1; \tag{3.35}$$

$$\frac{\partial}{\partial t}N(k,t) = \left[1 - a(t+\tau)^{-\alpha} - b\right]\frac{\partial}{\partial k}\left[N(k,t)G(k)\right] \Big/ \sum N(k',t)G(k') \quad \text{for } k > 1. \tag{3.36}$$

We now attempt to solve this system using the method of characteristics as in Sec. 2.2.3. However, we must first solve for $N(1,t)$ which will serve to fix the function along the characteristics. As in Sec. 2.2.3, we keep only the leading term in Eq. (3.35) and directly obtain the solution (using $N(1,1) = 1$):

$$N(1,t) \propto \frac{a}{1-\alpha}\left\{(t+\tau)^{1-\alpha} - (1+\tau)^{1-\alpha}\right\} + bt - b + 1\ . \tag{3.37}$$

The remaining equation is solved by multiplying both sides by $G(k)$; assuming that we are solving for the regime $k > k*$ such that $G(k) \sim k$ and $t > t*$ such that the normalizing sum is approximated by $\kappa(t+\tau)$; and considering (again) $Q \equiv Q(k(s), t(s) = k(s)N\left(k(s), t(s)\right)$:

$$\frac{\partial}{\partial t}Q + \frac{k\left[1-p(t)\right]}{\kappa(t+\tau)}\frac{\partial}{\partial k}Q = 0\ . \tag{3.38}$$

This equation is equivalent to $dQ/ds = (\partial Q/\partial t)(\partial t/\partial s) + (\partial Q/\partial k)(\partial k/\partial s) = 0$, from which we identify:

$$\frac{\partial t}{\partial s} = 1 \rightarrow t = s \tag{3.39}$$

and

$$\frac{\partial k}{\partial s} = k\frac{1-a(s+\tau)^{-\alpha}-b}{\kappa(s+\tau)} \rightarrow k = c_1 \mathrm{exp}\left\{\frac{1}{\kappa}\left[\frac{1}{\alpha}a\left(t+\tau\right)^{-\alpha} + (1-b)\ln\left(t+\tau\right)\right]\right\}\ . \tag{3.40}$$

Solving for Q as a function of that one initial condition, as we did in Sec. 2.2.3, we obtain in the limit where the exponential falls to one (as $t \gg 1$):

$$Q = \Psi\left(k(t+\tau)^{-(1-b)/\kappa}\right) \tag{3.41}$$

or for $N(k,t)$:

$$N(k,t) = \frac{1}{k}\Psi\left(k(t+\tau)^{-(1-b)/\kappa}\right)\ . \tag{3.42}$$

Comparing with the solution of Eq. (3.37)for $N(1,t)$ we get two regimes: if $b = 0$ the leading term of $N(1,t)$ is in $(t+\tau)^{1-\alpha}$ such that $\Psi(x) \sim x^{(1-\alpha)\kappa}$; if $b > 0$, the leading term is linear and $\Psi(x) \sim x^{\kappa/(1-b)}$. These scaling relations imply the following approximated scale exponents for the original distribution $N_k(t)$:

$$N_k(t) \sim \begin{cases} k^{-(1-\alpha)\kappa-1} & \text{if } b = 0 \\ k^{-\kappa/(1-b)-1} & \text{if } b > 0\ . \end{cases} \tag{3.43}$$

Note that if $\kappa = 1$, we fall back on the results of Simon's model if $b > 0$. We can thus have some confidence in this scale exponent [1]. However, the method of characteristics yields $\gamma = 1 + (1-\alpha)\kappa$ for the case $b = 0$; which is very different from the result of the main text $\gamma = 2 - \alpha\kappa$. We can easily test these result by choosing any growth function $G(k)$ leading to $\kappa \neq 1$ and birth function $p(t)$ with $b = 0$ and any given α. Figure 3.13 presents the distribution obtained after 10^7 iterations of the model with $G_{\mathrm{conv}}(k)$ ($\kappa = 3/2$) and $p(t) = t^{-2/3}$. The two methods predict $\gamma = 2 - \alpha\kappa = 1$ (main text) or $\gamma = 1 + (1-\alpha)\kappa = 3/2$ (method of characteristics). It is readily observed that the method of characteristics significantly overestimates the scale exponent.

[1] Note that in this case, our original asymptotic derivation implies a relation for α as a function of γ which we should not expect to be reversible. After an infinite time, the regime where α is relevant should be negligible in comparison to the (infinite) regime where $p(t) = b$.

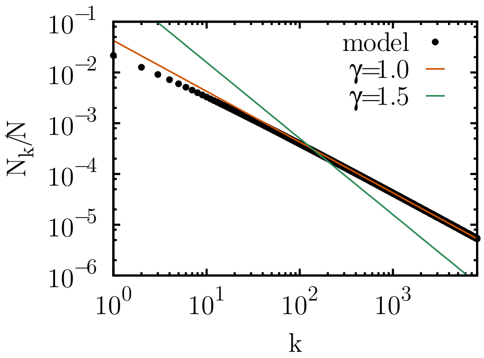


Figure 3.13 – **Verification of the scale exponent obtained with concave $G(k)$ and falling $p(t)$.** The model is iterated for 10^7 time steps and the two competing results for γ are compared to the obtained distribution.

Going back to our solution, it appears that the main problem could be in neglecting the second order term in the equation for $\partial N(1,t)/\partial t$. However, the same approximation was used in the case $p(t) \sim t^{-\alpha}$ in [Zanette & Montemurro, Journal of Quantitative Linguistics, 12:1 (2010)]. and lead to the correct result. Moreover, if $\kappa = 1$, we do recover the result of the main text ($\gamma = 2 - \alpha$). Perhaps re-injecting this approximated scaling behaviour in the main differential equation and considering higher order terms in the Kramers-Moyal expansion could lead to a better scaling estimate. Considering our initial asymptotic analysis of the rate equation led to the correct result, this approach has not been pursued.

Finally, we note that, if re-injected in the extremal criterion (see left-hand side relation in Eq. (8) of the main text), the results of this new analysis do respect the criterion while those obtained in the main text do so only if $\kappa = 1$. As we have validated our results in Fig. 3.13, this can only mean that the criterion itself is not exact. In fact, considering Fig. 3.12 we can postulate an explanation: the distribution $k^{-\gamma}$ is not a good approximation of the distribution, especially around $k_{\max}(t)$ and if $\kappa \neq 1$. It appears that the criterion gives us a good relation between derivatives, i.e. how do the position of $k_{\max}(t)$ changes in time in relation to how the population changes in time, but not a good approximation of the absolute position of $k_{\max}(t)$. Consider for instance how the right-hand side relation in Eq. (8) is respected for any power-law relation between $N(t)$ and $k_{\max}(t)$. Also supporting this hypothesis: using the results of Eq. (3.43) with the extremal criterion do not correctly reproduce the scaling exponent observed in Fig. 3.8 for $k_{\max}(t)$ while our first analysis of the rate equation does.

3.6 Appendix 3.C: Connection to network densification

This last section highlights the relation between the temporal scaling presented in this chapter and the so-called densification of complex networks. This last concept refers to the evolution of the ratio of links to nodes in connected systems. In the case of scale-free (or scale-independent) networks, this densification was observed to behave as a power-law relation between the number of nodes and the number of links [71]. Based on our previous results, we can conjecture a more precise relation.

In analogy to our theory, the number M of links would be directly proportional to the total number of events (or more precisely time $M = t/2$ as one link involves two nodes) while the number of nodes is directly related to the total population $N(t)$. Hence we expect the numbers of nodes and links to be related through the following expression

$$N(M) \simeq \frac{a}{\alpha+1}\left(2M+\tau\right)^{1-\alpha} - \frac{a}{1-\alpha}\tau^{1-\alpha} + 2bM. \tag{3.44}$$

With the usual $\alpha \leq 1$Eq. (3.44) can be rewritten as

$$N(M) \simeq \frac{a\tau^{1-\alpha}}{1-\alpha}\left(1+2M/\tau\right)^{1-\alpha} + 2bM - \frac{a}{1-\alpha}\tau^{1-\alpha} \tag{3.45}$$

to show that the relation is initially linear, i.e. when t and $M \ll \tau$,

$$\begin{aligned} N(M \ll \tau) &\simeq \frac{a\tau^{1-\alpha}}{1-\alpha}\left[1+(1-\alpha)\frac{2M}{\tau}+\mathcal{O}\left(\frac{M^2}{\tau^2}\right)\right] + 2bM - \frac{a}{1-\alpha}\tau^{1-\alpha} \\ &\simeq \left(2a\tau^{-\alpha}+2b\right)M. \end{aligned} \tag{3.46}$$

Equation (3.44) thus predicts an initially linear densification leading either to a second linear regime of different slope (b, steady state) or into a power-law relation if $b = 0$. This last behaviour is in fact observed in two network databases: the topology of Autonomous Systems (AS) of the Internet in interval of 785 days from November 8 1997 to January 2 2000 and the network of citations between U.S. patents as tallied by the National Bureau of Economic Research from 1963 to 1999 [71]. The results are presented on Fig. 3.14. Of the four systems considered in [71], these two were chosen to highlight two very different scenarios. On the one hand, the Internet can be reproduced by multiple pairs of a and τ parameters as long as $\tau \ll t$, since the system appears to have reached steady power-law behaviour. On the other hand, the patent citation networks do not fit with the power-law hypothesis as the system is transiting from a linear to a sub-linear power-law regime as $t \sim \tau$. This last scenario, while very different from a simple power-law growth as previously proposed, corresponds perfectly to the predictions of our theory.

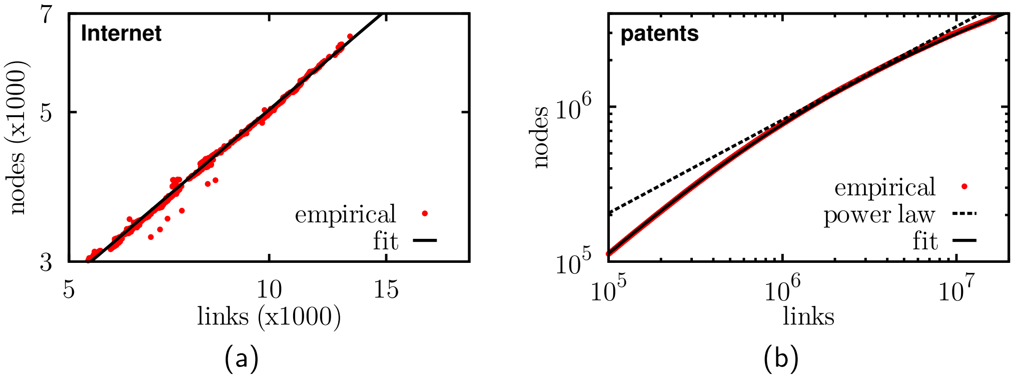


Figure 3.14 – **Densification of complex networks in log-log scale.** Densification, i.e. the relation between the number of nodes and the number of links, in two connected systems. (a) The Internet at the level of autonomous systems, reproduced by Eq. (3.44) with $a \simeq 1$, $\tau \simeq 0$, $\alpha = 0.16$ and $b = 0$. (b) The citation network of U.S. patents between 1963 to 1999, reproduced by Eq. (3.44) with $a = 5793$, $\tau = 875000$, $\alpha = 0.67$ and $b = 0$. The dotted line is the power-law relation claimed in [71].

Chapter 4

On growth II: Null models and their importance

Résumé

L'étude quantitative des systèmes sociaux complexes, des langages et des cultures humaines vit une effervescence remarquable depuis l'arrivée récente d'énormes bases de données. L'évolution du langage est un sujet particulièrement actif considérant que près de 6% des livres écrits ont maintenant été digitalisés, couvrant près de cinq siècles de développements culturels. Ce chapitre s'attaque au problème récent qu'est l'inférence des détails microscopiques d'un système à partir de ses propriétés macroscopiques. En particulier, nous traitons de la difficulté de distinguer, d'une part, la dynamique de notre découverte par échantillonage d'un système, et d'autre part, la dynamique de son évolution. La source de ce problème provient du couplage intrinsèque entre la structure des systèmes indépendants d'échelle et leurs propriétés temporelles. Pour illustrer ce phénomène, nous proposons de simples modèles nuls: un réservoir statique et des systèmes évoluant selon des règles fixes, mais avec un historique incomplet. Nous montrons ainsi comment la découverte progressive du premier modèle, ainsi que l'évolution des autres, reproduisent certaines propriétés dynamiques de vrais langages. Ces résultats démontrent comment des propriétés temporelles non triviales peuvent émerger simplement de notre connaissance incomplète du système à l'étude, et comment, même avec des données parfaites, les propriétés macroscopiques d'un système ne nous informent pas nécessairement sur les détails microscopiques de son évolution.

Summary

The study of complex social systems and culturomics, the quantitative studies of human languages and culture, has flourished in recent years with the advent of large datasets. The evolution of language has been a subject of particular interest considering that nearly 6% of books ever written have now been digitized, spanning over five centuries of cultural development. This chapter tackles the recent problem of inferring the microscopic details of a system based on its macroscopic properties. Namely, the difficulty of distinguishing between the dynamics of our discovery (or sampling) of a system and the dynamics of its actual evolution. The root of this problem stems from intrinsic coupling between the structure of scale-free systems and their temporal properties. To illustrate this, we propose simple null models: a static reservoir and systems growing with static rules with hidden past or incomplete data. We show how the progressive discovery of the former and the evolution of the others reproduce dynamical regularities of language. These results show how non-trivial temporal features can emerge merely from our incomplete knowledge of the system under study, and also how, even with perfect data sets, these features do not directly inform us about the rules governing their evolution.

4.1 Some theoretical processes and potential null models

We now present some potential null models of system evolution. By null models, we mean any processes whose rules can be simply stated and which can be used to compare to empirical data. The goal is to have some baselines against which we can test hypotheses. For instance, if a study observes macroscopic property A and concludes that A stems from a microscopic property B, we can compare the empirical data with results of null models which do not include B and see if A still occurs.

In fact, any growth process could be used as a null model in our context. However, we will only cover a few that include different assumptions. Our goal will be to show that all of these models feature the constraints uncovered in the last chapter — preferential attachment and temporal scale independence — either by design or as a consequence of their structure. These null models all lead to a scale-independent organization and to similar non-trivial temporal features regardless of their "microscopic" mechanisms. They can thus be used to test if certain macroscopic features observed in a recent culturomics study [87] can actually tell us anything about the microscopic details of language evolution.

For consistency with the following discussions, all of our null models will be presented as toy models for "language evolution" or, more accurately, for word usage. We already know that the overall distribution of word frequencies tend to roughly follow a power-law distribution, as per Zipf's law[1], with $\gamma < 2$. However, things are somewhat different when considering very large corpora to study the global evolution of language. In fact, very large data sets tend to feature two scaling regimes [46]. We will not focus on this feature too much as they can usually be reproduced by considering two populations (for function and content words).

Vocabulary size (the number of unique words) will be our main focus. When studying the growth of vocabulary size $N(K)$ as a function of the total number K of words written, one often comes across Heaps' law which states that the vocabulary size should roughly follow $N(K) \propto K^{\beta}$ with $\beta < 1$ [54]. While this is usually true for single text with $K \gg 1$, again different behaviours tend to appear when considering extremely small or extremely large corpora.

4.1.1 Random sampling

Our first null model of language evolution is neither a language (i.e. independent of context) nor evolving (i.e. static). Our model is inspired from the meta book concept of Bernhardsson *et al.* [16]. We suppose that written texts are essentially a random sampling from a bag of N_0 unique words, where a word i appears k_i times distributed following Zipf's law, i.e. the probability $P\left[k_i = k\right]$ is proportional to $k^{-\gamma_0}$. We use the subscript 0 for N_0 and γ_0 as the sample obtained from this bag will itself feature a vocabulary of size N and a potentially different scale exponent γ.

This model obviously follows the preferential attachment. Once a significant number of words have already been drawn, our best estimates for a given word's natural frequency is given by its frequency in our sample. This way, a word that has appeared x times is x/y times more likely to reappear next than a word that has appeared y times. This confirms our first constraint.

For our second constraint, we consider the probability $p(K)$ that the K-th written word is a new unique word (hence $p(K) = dN(K)/dK$). To calculate the probability $p(K)$ that the K-th word chosen from the reservoir is as of yet unseen, we merely sum over all words the probability that this word was never chosen in the $K-1$ prior tries and that it is now picked on the K-th try. Using $K_0 = N_0\langle k\rangle$, where $\langle k\rangle$ is the average number of occurrences for a given word in the reservoir, we write

$$p(K) = \sum_{i=\{\text{words}\}} \frac{k_i}{K_0}\left(1 - \frac{k_i}{K_0}\right)^{K-1} . \tag{4.1}$$

For $K \approx 1$, we easily find a linear growth ($N(K) \approx K$ as $p(K) \approx 1$)

$$p(K \approx 1) = \sum_{i=\{\text{words}\}} \frac{k_i}{K_0}\left(1 - \frac{k_i}{K_0}\right)^{K-1} \simeq \sum_{i=\{\text{words}\}} \frac{k_i}{K_0} = 1 . \tag{4.2}$$

[1] Zipf's law is simply the empirical observation of the power-law distribution of word occurrences in a text.

The behaviour for larger K is however more interesting. Using the fact that $k_i/K_0 \ll 1$ and $K/K_0 \ll 1$, since K_0 diverges when $N_0 \to \infty$ with $\gamma_0 < 2$ (typical value for Zipf's law in word occurrences), we can use the following equivalent form

$$p(K) \simeq \sum_{i=\{\text{words}\}} \frac{k_i}{K_0}\left(1-\frac{K}{K_0}\right)^{k_i} \ . \tag{4.3}$$

Or, transforming the sum over the number k of times a given word appears in the reservoir,

$$p(K) \simeq \sum_{k} \frac{kn_k}{K_0}\left(1-\frac{K}{K_0}\right)^{k} \ , \tag{4.4}$$

and using the scale-free distribution $n_k = Ak^{-\gamma}$ where A is a normalization constant,

$$p(K) \simeq \frac{A}{K_0}\sum_{k} k^{1-\gamma}\left(1-\frac{K}{K_0}\right)^{k} \ . \tag{4.5}$$

As $\gamma - 1 < 1$ and within the regime $K/K_0 \ll 1$, larger k values contribute significantly to the total sum such that it can be approximated by its integral

$$\begin{aligned} p(K) \simeq & \frac{A}{K_0}\int_0^\infty \left(1-\frac{K}{K_0}\right)^k k^{1-\gamma} dk \\ \simeq & -\frac{A}{K_0}\left[\ln\left(1-\frac{K}{K_0}\right)\right]^{\gamma-2}\Gamma\left(2-\gamma\right) \\ \propto & \left[\frac{K}{K_0} + \mathcal{O}\left(\frac{K}{K_0}\right)^2\right]^{\gamma-2} \propto (K)^{\gamma-2} \ . \end{aligned} \tag{4.6}$$

The vocabulary size

$$N(K) = \sum_{K'=0}^{K} p(K') \tag{4.7}$$

thus scales as

$$N(K \gg 1) \propto K^{\gamma-1} \tag{4.8}$$

in the regime $1 \ll K \ll K_0$ governed by the integral of Eq. (4.6).

4.1.2 Yule's process

As seen in Chap. 2, the Yule process is one of the first model for the growth of scale-independent systems. As with our first null model, the Yule process produces a context free system (or language); but one which evolves, albeit with static rules. Let us recall the process, originally a mathematical model of evolution. Each species (every written word) in the ecosystem (text) undergoes specific mutations which create a new species (a new word is written) of the same genus (unique word already in vocabulary) at a rate s; and each genus (unique word in the vocabulary) undergoes generic mutations which create new species (a new word is written) of a new genus (unique word new to the vocabulary) at a rate g. Basically, in our context, words are written in a text of K words with a vocabulary of size N at a rate $sK + gN$ where the rate sK corresponds to occurrences of existing words and gN to the introduction of new unique words. This process leads to a distribution of words following Zipf's law, i.e. word occurrences following $P\left[k_i = k\right] \propto k^{-\gamma}$ with $\gamma = 1 + g/s$.

The Yule process obviously follows preferential attachment by design; which is the first of our two constraints. As for the vocabulary of the Yule process, we know that in continuous time t it follows an exponential growth $N(t) = e^{gt}$ and we are interested in translating that as a function of the total number of written words K. As the number of occurrences of a unique word who first appeared at time t' follows an exponential growth as $e^{s(t-t')}$, we can obtain the total number of written words $K(t)$ by integrating these occurrences over the number of words that appeared during every infinitesimal step dt'. We write

$$K(t) = e^{st} + \int_0^{t'} \dot{N}(t')e^{s(t-t')}dt' = e^{st} + \frac{g}{g-s}e^{st}\left[e^{(g-s)t} - 1\right] \ , \tag{4.9}$$

where the first term corresponds to the original word. Since the distribution of written text always follow a scale exponent $\gamma < 2$, we can use the fact that $s > g$ and $t \gg 1$ to eliminate the second exponential and approximate $K(t)$ as

$$K(t) \simeq e^{st}\left(1 - \lambda\right) \tag{4.10}$$

with $\lambda = g/(g-s)$. We can thus obtain time as a function of the number of written words,

$$t = \frac{1}{s}\left[\ln K - \ln\left(1-\lambda\right)\right] \tag{4.11}$$

such that we can write the vocabulary growth $N(K)$ as

$$N(K) = \exp\left\{\frac{g}{s}\left[\ln K - \ln\left(1-\lambda\right)\right]\right\} \propto K^{g/s} \propto K^{\gamma-1} \ , \tag{4.12}$$

which is exactly the dominant behaviour observed in the case of the random sampling.

4.1.3 Incomplete history of preferential attachment

We now consider a null model that encompasses both the incomplete knowledge of random sampling and the static growth rules of Yule's process: a system (language) grows according to preferential attachment and we start reading its output only after the first K_0 events. Effectively, this emulates how we often do not witness (or gather data on) the infancy of a system (be it a language, the Internet or a social system). This null model will help us quantify the impact of the finite size in the sampled reservoir.

This "incomplete history of preferential attachment" is essentially equivalent to a combination between a preferential attachment growth (when new words are used after K_0) and sampling of a static reservoir (when words created before K_0 are reused after K_0). As per Eq. (3.17), we consider two cases for the underlying PA process: a decreasing birth rate $\tilde{p}(K) \propto K^{\gamma-2}$ or the limit $\tilde{p}(K) \to 0$. The tilde is here used to identify properties of the underlying hidden PA process.

Once again, as both the underlying process and the sampling of the unknown history follow preferential attachment, we assume that $G(k) \propto k$ (at least for large values of k) is respected. Similarly, the birth rate $p(K)$ can also be expected to follow its usual constraint as per the two processes that we are here combining. We can write, for all possible $\tilde{p}(K)$ and K_0,

$$p(K) = \tilde{p}(K) + (1 - \tilde{p}(K)) \left(\frac{K_0}{K_0 + K} \right) \hat{p} \left(\hat{K}(K) \right) \tag{4.13}$$

where time K is for the incomplete reader (i.e. $K = 0$ corresponds to $\tilde{K} = K_0$ for the underlying process), $\hat{p}(x)$ is the expected "birth rate" in our random sampling model of Sec. 4.1.1, and $\hat{K}(K)$ corresponds to the average number of words that have been drawn from the reservoir at time K. The term proportional to $(1 - \tilde{p}(K))$ in Eq. (4.13) is essentially the probability that the underlying process selects an old word $[(1 - \tilde{p}(K))]$ times the probability that this old word is picked from the reservoir of size K_0 and not from the new observed reservoir of size $[K_0/(K_0 + K)]$ times the probability that the word picked from the reservoir is an occurrence of a word never seen before by the incomplete reader $[\hat{p}(\hat{K}(K))]$. This average $\hat{K}(K)$ is given by the number of past events which picked words from the reservoir. It can be written as

$$\hat{K}(K) = \sum_{i=0}^{K-2} (1 - \tilde{p}(K)) \left(1 - \frac{i}{K_0 + i} \right) . \tag{4.14}$$

Looking only at qualitative behaviour, we can use Eq. (4.6) which yields

$$p(K) \simeq \tilde{p}(K) + (1 - \tilde{p}(K)) \left(\frac{K_0}{K_0 + K} \right) \left[\sum_{i=0}^{K-2} (1 - \tilde{p}(K)) \left(1 - \frac{i}{K_0 + i} \right) \right]^{\gamma-2} . \tag{4.15}$$

In the regime $K_0 \gg K$. Equation (4.15) becomes

$$p(K) \simeq \tilde{p}(K) + (1 - \tilde{p}(K))^2 \left[\sum_{i=0}^{K-2} \left(1 - \frac{i}{K_0} \right) \right]^{\gamma-2} . \tag{4.16}$$

The sum can be evaluated straightforwardly

$$p(K) \simeq \tilde{p}(K) + (1 - \tilde{p}(K))^2 \left[\frac{K-1}{K_0} \left(K_0 - \frac{K-2}{2} \right) \right]^{\gamma-2} \tag{4.17}$$

and again using $K_0 \gg K \gg 1$,

$$p(K) \simeq \tilde{p}(K) + (1 - \tilde{p}(K))^2 \, K^{\gamma-2} , \tag{4.18}$$

we obtain the following leading behaviours for the two general cases of $\gamma < 2$ and $\gamma \approx 2$ which can be obtained using Eq. (3.17) with $b = 0$ or with $\alpha = 2 - \gamma$ and $b \approx 0$ respectively:

$$p(K) \propto \begin{cases} K^{\gamma-2} + \left(1 - K^{\gamma-2}\right)^2 K^{\gamma-2} \sim K^{\gamma-2} & \text{if } \tilde{p}(K) \sim K^{\gamma-2} \\ K^{\gamma-2} & \text{if } \tilde{p}(K) \sim 0 \, . \end{cases} \tag{4.19}$$

In the regime $K \approx K_0 \gg 1$ and beyond. The finite sum in Eq. (4.15) can be evaluated by applying the Euler-Maclaurin formula [1]. Keeping only the leading term, which is equivalent to directly approximating the sum by an integral, we get

$$p(K) \simeq \tilde{p}(K) + \left(1 - \tilde{p}(K)\right)^2 \frac{K_0}{K_0 + K} \left[K_0 \ln\left(K_0 + K\right)\right]^{\gamma-2} \tag{4.20}$$

whose leading behaviours are

$$p(K) \propto \begin{cases} K^{\gamma-2} + \left(K_0 + K\right)^{-1} \left(1 - K^{\gamma-2}\right)^2 \left[\ln\left(K_0 + K\right)\right]^{\gamma-2} \sim K^{\gamma-2} & \text{if } \tilde{p}(K) \sim K^{\gamma-2} \\ \left(K_0 + K\right)^{-1} \left[\ln\left(K_0 + K\right)\right]^{\gamma-2} \sim K^{-1} & \text{if } \tilde{p}(K) = b \sim 0 \, . \end{cases} \tag{4.21}$$

The last case gives us an idea of how the finite size of a reservoir affects the "vocabulary growth" for large K, either here or in the model of Sec. 4.1.1. Integrating $p(K) \sim K^{-1}$ over an interval of very large K leads to a logarithmic growth, a limiting case already observed in the previous chapter for the pathological case $\gamma = 2$.

4.2 On the problem of distinguishing evolution from discovery in language and other scale-independent systems.

Linguistics, the study of language, is a science of the microscopic. Words form a language only when put in structure, form, and context. The recent creation of a large corpus of written texts by the Google Books Team opens the door for a different approach using large datasets and a quantitative methodology. However, in considering words as mere data points, one risks neglecting the structure, form, and context of words, albeit at the benefit of painting a broader picture through macroscopic statistical features. It is in this context that statistical regularities of written text and of language evolution have been identified: Zipf's and Heaps' laws, and more recently cooling patterns following language expansion [87]. These patterns show a power-law convergence for the observed frequencies of unique words through time, as well as a power-law reduction of the rate at which new words are introduced. It is yet unclear how to reconcile the macroscopic results of theses studies with the microscopic nature of language.

In this section, we focus on the aforementioned statistical regularities and their potential relevance (or lack thereof) to the mechanism behind language evolution. Considering that, although impressively large, the available corpora constitute only samplings of a given language at a given time, there must exist certain dynamical properties inherent to the dynamics of having a progressively better sampling of a complex system as the years progress (i.e. artefacts of the methodology). How can we distinguish the sampling dynamics from the actual evolution of the language under study? Without this distinction, it is hazardous to infer anything about human languages from macroscopic properties of the corpora. Moreover, even with a perfect sampling, it is unclear how these macroscopic temporal features of language are even related to its temporal evolution since the structure of the language can itself affect certain temporal variables (e.g. vocabulary size and any finite size fluctuations). How can we then distinguish the effects of language structure from actual language evolution?

To answer the first question, we use our simplest null model of language evolution: the random sampling of a scale-independent system. We now also impose an ad hoc “syntactic” constraint that a small number of words must appear on average once every s_0 words. This last constraint is enforced to emulate the role of articles and other highly frequent function (or structure) words. We use only one such word since the exact number is insignificant as long as the total frequency s_0 is respected. The constraint is introduced to differentiate families of language, as for instance, Indo-European languages (like English) feature a higher ratio of function to content words (high s_0) than most Asian languages.

To answer our second question, we use the first model of evolution for scale-independent systems: the Yule process. As with our first null model, the Yule process also produces context free language; but one which will evolve with static rules. We enforce two additional constraints: the aforementioned syntactic rule and an initial unobserved basin of words (species) of size N_0. This basin represents an initial set of words, but is unobserved in the sense that these words are still unknown until their first specific mutation.

4.2.1 “Evolution” of vocabulary growth in languages

The methodology of our study is simple. We chose two languages with significantly different behaviours, Spanish and Chinese, that reproduce the statistical features of most other languages. We produce null models with different set of parameters to emulate these two languages. Both the languages and the null models were then analysed in the same fashion. For a given year t with a corpus of $K(t)$ written words, we computed the vocabulary size $N(K)$ in both the languages and a corpus of size $K(t)$ built from their corresponding null models.

As mentioned in introduction, Heaps’ law usually governs $N(K)$ as it roughly follows $N(K) \propto K^{\beta}$ with $\beta < 1$ for single text with $K \gg 1$. However, different behaviours tend to appear when considering small or extremely large corpora and one can then expect to see more than

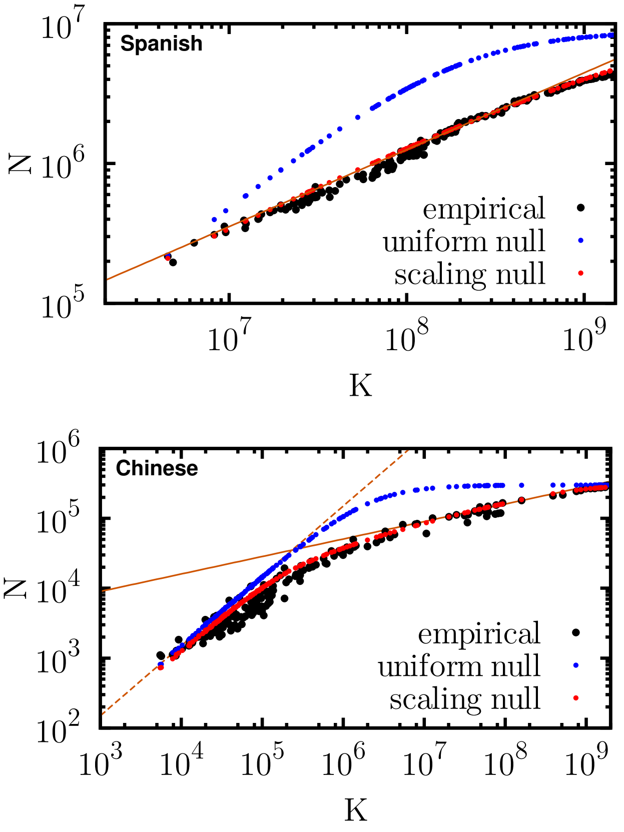


Figure 4.1 – **Vocabulary size as a function of corpus size versus a random sampling.** Number of unique words found in corpora — real or sampled from the null models — of different sizes. The results are based on (top) Spanish and (bottom) Chinese corpora. The null models are tuned to best reproduce the empirical results and use the following parameters. (Spanish) Null model with scaling: $\gamma = 1.5$, $N_0 = 9 \times 10^6$ and $s_0 = 0.5$; Null model with uniform distribution: $N_0 = 9 \times 10^6$ and $s_0 = 0.95$. (Chinese) Null model with scaling: $\gamma = 1.25$, $N_0 = 3 \times 10^5$ and $s_0 = 0.3$; Null model with uniform distribution: $N_0 = 3 \times 10^5$ and $s_0 = 0.85$. The full orange lines correspond to scaling with exponent $\gamma - 1$ and the dotted orange line is a simple linear scaling. Also, the Spanish vocabulary growth give us a good example of the effects of the reservoir's finite size in our null model. This effect, according to Eq. (4.21), is a transition from a scaling behaviour to a logarithmic growth.

one scaling regions with progressively lower β. In a recent study [87], this last feature was described as a result of a diminishing need for new words; hence the apparent slow-down (or "cooling") of vocabulary growth. However, a similar behaviour can be observed in our null models (see Fig. 4.1). In this case, mostly as a consequence of the scale-independent distribution of word frequencies. This is illustrated with an equivalent sampling model based on an underlying uniform distribution which fails to reproduce the scaling regime: $N(K)$ then merely scales linearly with finite size effects for very large K.

While it is appealing to try and use the huge data sets of the Google Books project to study the underlying mechanisms of language evolution, our results provide an important warning: macroscopic data are not necessarily a good proxy for microscopic events and properties, no

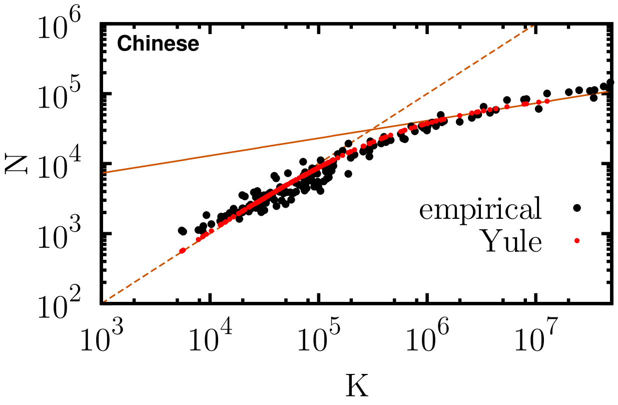


Figure 4.2 – **Chinese vocabulary size versus Yule's process.** Yule's process uses mutation rates $g = 1$ and $s = 4$, an initial basin of size $N_0 = 150000$ and a syntactic constraint $s_0 = 0.9$. Note that this syntactic constraint is much less realistic than those needed to fit the sampling model to the data. As before, the dotted orange line corresponds to a linear scaling and the full orange line to a scaling with exponent $\gamma - 1 = g/s = 0.25$. This double scaling behaviour was claimed in [87], and while the empirical data might not provide much support for this particular hypothesis, our analyses of both Yule's process and the of the random sampling model demonstrate the presence of two scaling regimes.

matter how many of those are compiled. For instance, logic might dictate that the initial linear regime observed in the Chinese dataset is most likely a sampling effect and not a fast (linear) language expansion. In fact, it is interesting that the syntactic constraint ($s_0 = 0.3$) used in the null model to reproduce the vocabulary growth is a good estimate of the actual frequency of function words in Chinese, which is thought to be around 0.25 [25]). We can expect the same to be true for Spanish where we used $s_0 = 0.5$, as that fraction is almost 0.6 in the English language [86], and both are Indo-European languages with similar syntax. That being said, the observed behaviour could also be caused by an evolution starting with some initial population (as in Yule's process, see Fig. 4.2) or by an era with a significant "need for new words" as postulated in [87]. At this point, it might important to stress that the two power-law regimes assumed in Figs. 4.1 and 4.2 are tenuous as other curves might reproduce might also reproduce the observed behaviour. However, we follow the methodology of the publication we are using as a case study (i.e., Ref. [87]) and while the empirical data might not be convincing, our analytical results do show that our null models do feature a double scaling behaviour.

4.2.2 Cooling patterns in language "evolution"

While the previous results concerned the evolution of two properties through time, namely the occurrence distribution and vocabulary size, this section focuses on a more dynamical feature of written text: standard deviation of growth rate fluctuations $\sigma(K)$. This quantity is more dynamical in the sense that it depends not only on the state of the system during year

t, but also on its state during the previous year, i.e. $t-1$. To compute $\sigma(K)$, we must first compute the growth rate of each unique word between year $t-1$ and year t. To this end, we assume an exponential growth such that the growth rate $r_i(t)$ of word i is given by:

$$r_i(t) = \log\left(k_i(t)\right) - \log\left(k_i(t-1)\right) \ . \tag{4.22}$$

Using only a fraction of significantly frequent words (we use the same frequency cut-off[2] as in the previous study [87]), $\sigma(K)$ is then computed from this ensemble of $r_i(t)$.

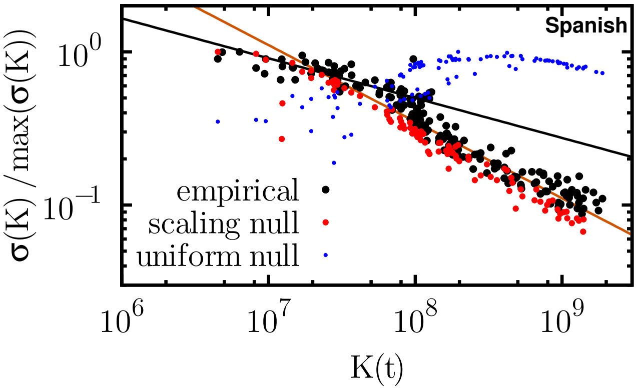
In this particular case, we only use the random sampling as our sole null model. The point we wish to make is that the "cool down" observed in Ref. [87] can be mostly explained by language being a random sampling of a non-evolving (or static) reservoir of words. While the null model itself is obviously wrong, sampling dynamics could still be the main driving force behind some temporal patterns observed in the Google Books corpora. This is confirmed in Fig. 4.3, where the sampling behaviour produces a qualitatively similar convergence of $\sigma(K)$ as observed in the empirical data. Importantly, we should note that our results do not reproduce the cooling exponents observed in Ref. [87]. The reasons for this discrepancy is unfortunately unknown. The previous study defines corpus size as the number of written words belonging to words above the cut-off, whereas we define corpus size as the total number of written words. However, we used both definitions and we can state that this is not the source of the difference in observed cooling exponents. It may also be caused by an update in the data set (we use the version published on July 1st 2012). At the very least, we know that both our two data sets (Chinese and Spanish corpora) and our null models were analysed in exactly the same fashion.

4.2.3 Discussions (and why preferential attachment is not a mechanistic assumption)

The fact that the study of our null models reproduce certain features of the study of actual human languages does not tell us anything about the evolution of the languages, and certainly does not imply that there is a mechanistic link between the models and reality.

What we can conclude is that the random sampling of a static scale-independent system can itself feature scaling behaviour on diverse fronts, namely in vocabulary growth and the cooling patterns of the growth fluctuations. Thus, patterns of temporal scaling observed in culturomics stem not necessarily from dynamical properties of languages, but perhaps merely from temporal properties of the observation method (e.g. sampling) or from our incomplete knowledge (hidden past), both of which are closely coupled to the organisational properties of the sampled system. It thus appears useless to infer microscopic mechanisms

[2] Words are kept if their occurrence frequency is at least $10/\min\{K(t)\}$, meaning that they would appear on average at least 10 times even in the smallest corpus available.



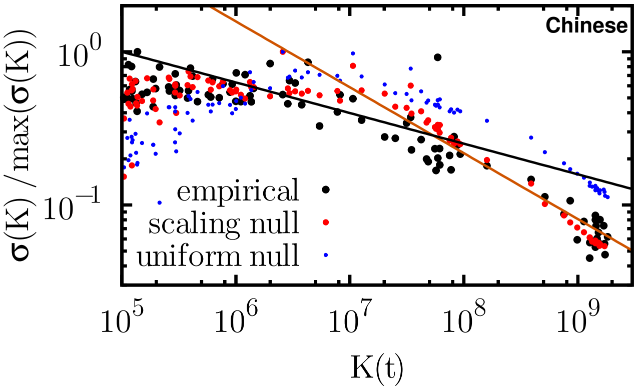


Figure 4.3 – **Standard deviation in word usage growth rate as a function of corpus size.** Standard deviation of growth rates $r_i(t)$ found in subsequent corpora — real or sampled from the null models — of different sizes. The results are based on (top) Spanish and (bottom) Chinese corpora and the null models use the same parameters as for Fig. 4.1. The null model with scaling follows the correct cooling pattern whereas the uniform null model only approximately does so in one case. Note that the null models always end up scaling as $\sigma(K) \propto K^{-1/2}$ (orange line) which is a standard sampling behaviour. The black line follows the scaling originally observed in Ref. [87]

from macroscopic temporal properties as those are often universal consequences of the scale-independent organisation of language and culture.

On the other hand, there is no denying that human languages have evolved over the last centuries, and the models presented here have no pretension beyond that of being simple null hypotheses. The main point is that coarse-grained studies of language evolution are, at least in regards to the methodology presented here, not able to differentiate between actual evolving human languages and our sampling model. In fact, this is not a surprising conclusion as the effects of evolution are mostly microscopic: certain words become more or less frequent as time goes by, some will die and others will emerge, but perhaps in a self-averaging way. It then goes without saying that this phenomenon should be studied through microscopic means, e.g. the evolution in the usage of unique words, rather than by a macroscopic statistical approach.

The same conclusions hold for any scale-independent system. The recipe for models of scale independence is quite simple: the constraints we have discussed up to now are necessary ingredients and other mechanistic assumptions are mostly for flavour. From there, most macroscopic features are influenced by the scale-free organization of the model and not necessarily direct consequence of the mechanistic assumptions.

As we have seen in our null models, even preferential attachment does not tell us anything about the underlying mechanisms. In fact, while we use preferential attachment has a literal mechanism[3], it can also be an artefact (or symptom) of other growth mechanisms. For instance, in percolation on a lattice, it can be shown that the relation between the growth rate of a cluster and its size is asymptotically linear (see Sec. 4.3). This is due to the asymptotically linear relation between a cluster's size and its parameter, but does not imply that sites are given to clusters because of a preferential attachment process. Similar analogies can be made for social sciences where preferential attachment might merely captures how good an individual is at a given activity, but does not imply that one is good *because* of one's past activity.

This is not so much a shortcoming of our framework as statistical physics usually describes systems in a macroscopic manner without the corresponding microscopic details. Consider how the canonical ensemble does not depend on how a given system is coupled to a heat bath (or energy reservoir), but simply on the fact that this coupling exists. For our framework to be universal, we have to accept to lose some level of specificity. The objective of this section was to both show the limits of our own work, and serve as a warning regarding inference of social or cultural mechanisms from macroscopic models.

4.3 Appendix 4.A: Percolation and self-organized criticality

In Chap. 1 we rapidly presented a few examples of scale-independent systems and we now highlight a link between two of those: earthquakes and percolation. If you recall, earthquake magnitudes followed a power-law distribution such that even though most earthquakes are of negligible magnitude, a microscopic number of them still hold a macroscopic fraction of all energy released by earthquakes. We then introduced percolation as a toy model to investigate the link between power-law distribution and this organization apparently in-between homogeneous and heterogeneous states. In so doing, we clarified the concept of criticality: this phase transition occurs in percolation only around a well-defined value. We then asked the following question: Does this imply that the physics of earthquakes is also tuned to a precise phase transition? Just like the null models presented in this chapter, percolation and its variants can be used to study real systems, perhaps not languages, but systems which are

[3] Meaning that in Chap. 3, it is literally because an individual has share k that he gets $G(k) \sim k$ chances to get richer.

more physical in nature, like earthquakes. They can thus be used to tackle questions like the one presented above.

In light of the growth constraints introduced in the previous chapter, and observed in the previous null models, we can ask whether or not percolation follows the preferential attachment principle and temporal scaling in the introduction of new individuals (clusters). We could assume that a larger cluster is in fact more likely to gain new occupied sites from its perimeter than a smaller cluster, such that preferential attachment would be respected. However, percolation systematically introduce new resources (occupied sites) and has a finite limit to its population size (lattice dimensions). Without control over $p(t)$, the system can overshoot its phase transition and simply grow toward a fully clustered (heterogeneous) state. This is why the occupation probability u is so important in percolation, it grants us some control over the total resource $K(t)$.

Therefore, we now wish to confirm that preferential attachment is respected at least for large clusters, and introduce an adaptive control of the system over its own $p(t)$ to remove the need for critical tuning of the occupation probability.

4.3.1 Preferential attachment and temporal scaling in percolation

We want to show that, when percolating sites are occupied randomly (in a sequential manner), clusters of size x are x/y times more likely to grow with the next occupied site than clusters of size y (at least for $x \gg 1$ and $y \gg 1$).

Let us consider any geometry where the number of neighbouring sites of a cluster tends to grow with their size. This assumption is not limiting at all as it covers all interesting cases (except some pathological cases such as percolation on a one-dimensional chain as considered in Chap. 1). Let us also consider percolation as a dynamical process: at each time step, a site is selected for occupation and is occupied with probability u and left empty for all time with probability $1 - u$.

We define the perimeter l of a cluster as the number of non-occupied neighbouring sites which may or may not have been tested for occupation yet. Suppose that we know exactly the numbers N_{kl} of possible clusters of size k and perimeter l. By *possible*, we mean that we are considering a simple enumeration of possible cluster structure independently of u. With $\{N_{kl}\}$, we can write the size distribution $n_k(u)$ like so [43]

$$n_k(u) = u^k \left[\sum_l N_{kl} \bar{u}^u \right] \tag{4.23}$$

where we ask for k sites to be occupied (probability u^k) and for the perimeter of size l to be unoccupied (probability $\bar{u}^l$ with $\bar{u} = 1 - u$). Similarly, we can write the distribution of perimeter given a size k, noted $n_l^{(k)}(u)$, as

$$n_l^{(k)}(u) = \frac{N_{kl}\bar{u}^l}{\sum_{l'} N_{kl'}\bar{u}^{l'}} \ . \tag{4.24}$$

With this new distribution, we can calculate the mean perimeter $\mu_k(u)$ of clusters of size k:

$$\mu_k(u) = \frac{\sum_l lN_{kl}\bar{u}^l}{\sum_l N_{kl}\bar{u}^l} = \bar{u}\frac{d}{d\bar{u}}\log\left[u^{-k}n_k(u)\right] \tag{4.25}$$

where the second equality comes from isolating the sum in Eq. (4.23). We now have an expression for the average perimeter of clusters of size k which depends only on u and on the distribution of cluster size. We are now interested in the behaviour of this expression in the limit $k \to \infty$ knowing that preferential attachment would imply $\mu_k(u) \propto k$. We are thus wondering whether $k^{-1}\mu_k(u)$ goes to a constant in the following limit,

$$\lim_{k\to\infty} k^{-1}\mu_k(u) = \lim_{k\to\infty} k^{-1}\bar{u}\frac{d}{d\bar{u}}\log\left[u^{-k}n_k(u)\right] \ . \tag{4.26}$$

We first find

$$\lim_{k\to\infty} k^{-1}\log\left[u^{-k}n_k(u\right] = \lim_{k\to\infty}\left[k^{-1}\left(-k\log u\right) + k^{-1}\log\left(n_k(u)\right)\right] \tag{4.27}$$

where the second term necessarily goes to zero (product of two decreasing terms), such that

$$\lim_{k\to\infty} k^{-1}\log\left[u^{-k}n_k(u)\right] = -\log u \ . \tag{4.28}$$

With this result, we finally find that percolation does in fact follow the preferential attachment principle,

$$\lim_{k\to\infty} k^{-1}\mu_k(u) = \bar{u}\frac{d}{d\bar{u}}\left[-\log u\right] = \frac{\bar{u}}{u} \ , \tag{4.29}$$

meaning $G(k) \propto \bar{u}k/u$ for $k \to \infty$.

Preferential attachment is one of two constraints that we want to see respected in models of scale-independent growth. The second is temporal scaling. We thus check how the cluster population size $N(t)$ scales with regard to the number of occupied sites t in classical percolation. Hence, using the number of occupied sites as a measure of time allows us to consider percolation as a dynamical process. More precisely, we consider a two-dimensional square lattice where at each step ($i \to i+1$) one of L sites is randomly chosen for a percolation trial. With probability u, the site is occupied and the system clock is increased by one ($t \to t+1$). Sites can only be chosen once for this trial.

We study the behaviour of connected clusters. What is the probability $p(t)$ that the t-th occupied site creates a new cluster of size one (none of its four neighbours is occupied yet)? This probability defines the ratio between the growth of the cluster population to the number of occupied sites. We want to test whether or not $p(t)$ will follow our functional form.

The t-th occupied site will mark the birth of a new cluster if none of his four neighbours were among the $(t-1)$ first occupied sites. We can then directly write

$$p(t) = \prod_{j=1}^{t-1} \left(1 - \frac{4}{L-j}\right) \ . \tag{4.30}$$

Rewriting $p(t)$ as

$$p(t) = \prod_{j=1}^{t-1} \frac{L-j-4}{L-j} \tag{4.31}$$

one can see that

$$p(t) = \prod_{j=1}^{4} \frac{L-t+1-j}{L-j} \qquad \text{for } t > 4. \tag{4.32}$$

For $L \gg t \gg 1$,

$$p(t) \simeq \frac{(L-t)^4}{L^4} = \frac{(t-L)^4}{L^4} \ . \tag{4.33}$$

Equation (4.33) agrees with Eq. (3.17) using $a = L^{-4}$, $\tau = -L$, $\alpha = -4$ and $b = 0$, see Fig. 4.4.

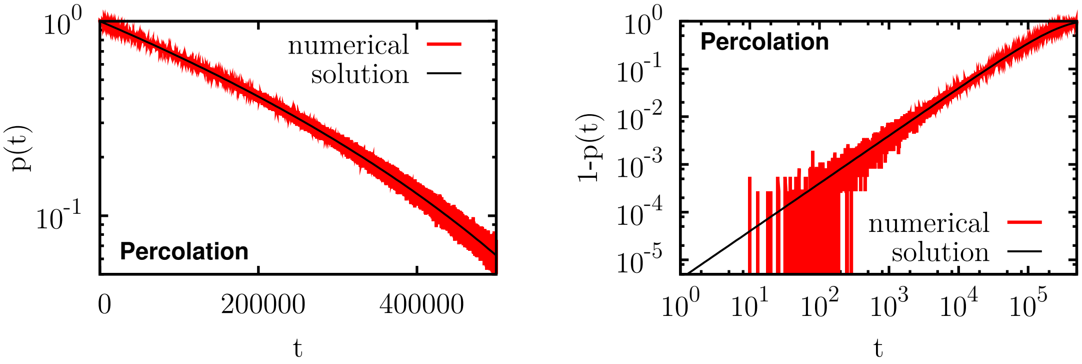


Figure 4.4 – **Percolation on a 1000x1000 square lattice at phase transition ($\mathbf{u_c = 0.5927\ldots}$).** (left) The evolution of the probability $p(t)$ that the t-th occupied site results in the creation of a new cluster (in semi-log plot). (right) Log-log plot of the complementary probability $1 - p(t)$ to highlight the initial temporal scaling. The solution corresponds to Eq. (4.33).

It is important to note that Eq. (4.33) does not depend on the percolation probability u. Under this form, the ability of this system to converge towards its critical state depends on the number of sites occupied, i.e. time t. Noting that $t \equiv uL$, the critical time $t_c = u_c L$, corresponding to the critical point in u, could perhaps be calculated through a self-consistent argument on the number of occupied sites required by the scale-free distribution of cluster size in the critical state. This is however, not our current concern.

4.3.2 Bak-Tang-Wiesenfeld model

We want a model that is loosely based on percolation, and will thus inherit its preferential attachment property, but that does not need an external tuning of the population size

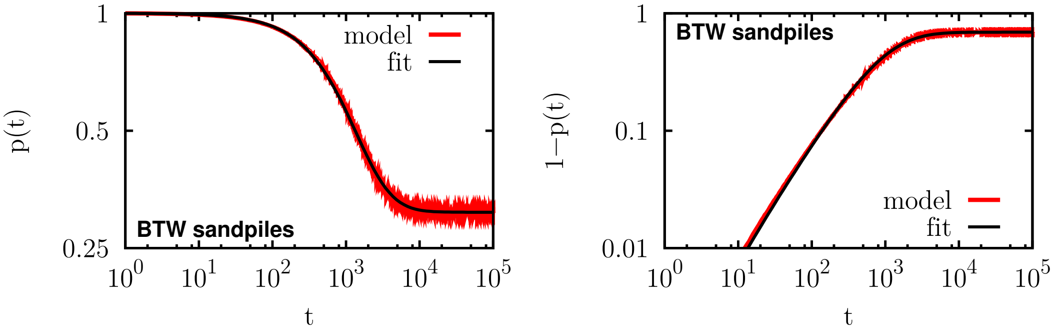


Figure 4.5 – **Cluster population in the Bak-Tang-Wiesenfeld model.** We use a **64**x**64** square lattice with critical height $\mathbf{z^* = 4}$. (left) The evolution of the probability $p(t)$ that the t-th site to reach a height of $z^* - 1$ results in the creation of a new potential avalanche (in log-log plot). (right) Plot of the complementary probability $1 - p(t)$ for growth events. The fit uses Eq. (3.17) with $b = 0.3090$, $\alpha = 3.5$, $\tau = 3000$ and a fixed by $p(1) = 1$ (i.e. $a = (1 - b)(1 + \tau)^{\alpha}$).

through the percolation probability. The first model of self-organized criticality, the Bak-Tang-Wiesenfeld process (BTW), shares a similar basic mechanism with percolation and, interestingly, was introduced as a thought experiment on the distributions of energy released in avalanches and earthquakes [11]. Their model follows the evolution of a sandpile where grains of sand are randomly dropped on a two dimensional square lattice. The sand tower on the site (i, j) crumbles on its four nearest neighbours when its height $z_{i,j}$ reaches a critical height $z^* = 4$. The model can thus be followed by iterating the following algorithm:

1. *Initialisation.* Prepare the system in a stable configuration: we choose $z_i = 0 \; \forall \; i$.
2. *Drive.* Add a grain at random site i.
$$z_i \rightarrow z_i + 1 \; .$$
3. *Relaxation.* If $z_i \geq z^*$, relax site i and increment its 4 nearest-neighbours (nn).
$$\begin{aligned} z_i &\rightarrow z_i - 4 \; , \\ z_{nn} &\rightarrow z_{nn} + 1 \; , \end{aligned}$$
Continue relaxing sites until $z_i < z^*$ for all i.
4. *Iteration.* Return to 2.

When this algorithm reaches its steady-state, the distribution of cluster sizes at any given time follows a power law with a cut-off (upper bound) determined by the size of the lattice. Interestingly, the relation between this upper bound and the lattice size is of the same form as the relation between the peloton position and time in preferential attachment processes (see Sec. 2.2.1.2 and Ref. [26] §3.9).

Figure 4.5 follows the BTW sandpile model on a 64x64 square lattice and we can see, almost surprisingly, that the delayed temporal scaling function offers a good approximation of the birth rate of critical clusters through time [4]. More precisely, we followed the probability $p(t)$

[4] Critical clusters are defined as clusters of connecting sites with height $z^* - 1$.

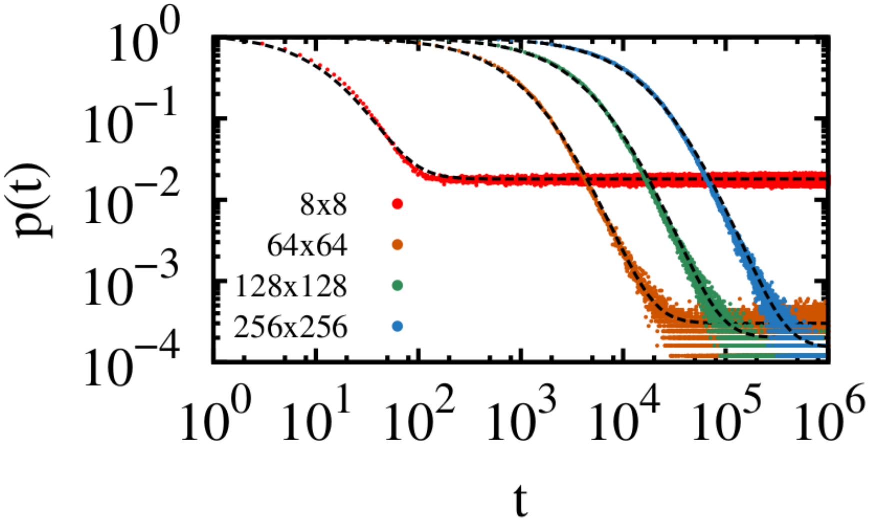


Figure 4.6 – **More cluster populations in the Bak-Tang-Wiesenfeld model.** We now follow the evolution of the probability $p(t)$ that the t-th site to reach a height of $z^* - 1$ because of a manually inserted grain of sand results in the creation of a new potential avalanche (critical cluster). Different size L of square lattice (LxL) are investigated: 8x8 leads to $b \simeq 0.018$, 64x64 to $b \simeq 0.0003$, 128x128 to $b \simeq 0.0002$, and 256x256 to $b \simeq 0.00015$. All are fitted with $p(t) = a(t + L^2/2)^{-3.5} + b$ with a_i fixed to force $p(1) = 1$.

that the t-th site to reach a height of $z^* - 1$ marked the birth of a new critical cluster, this site only had neighbours with height $< z^* - 1$. The temporal scaling exponent is also very close to the one observed in percolation, 3.5 instead of 4.0, and both are significantly greater than those observed in social or man-made systems. Similarly, the steady-state value, $b \simeq 0.3$ is far greater than those previously observed.

To change the birth rate parameters observed in the BTW model, we can adjust our metric, lattice size, or the geometry. Changing our metric can mean, for instance, looking only at sites that reach a height of $z^* - 1$ when we "manually" drop a grain of sand and ignore all events caused by self-organization (avalanches). This is done in Fig. 4.6 and leads to a much lower value for the steady-state b. We can also change lattice size and affect both b and the temporal delay τ. In fact, as seen in Fig. 4.6, we roughly observe $\tau \simeq L^2/2$ on a LxL lattice. While we have yet to do so, we can assume that changing the geometry of the lattice would affect all parameters, including the temporal scaling exponent α. Thus, controlling the lattice geometry and size, we could tune the BTW model to reproduce social systems with an asymptotic steady-state. While this was not done, the idea of using a sandpile process to mimic a sexual system or the productivity of artists and scientists is certainly appealing. Moreover, the death of critical clusters who cause avalanches opens the door to a generalization of our framework to include removal of individuals.

Chapter 5

On structure I:
Networks beyond preferential attachment

Résumé

Nous démontrons maintenant comment l'introduction de *balles colorées* dans le problème d'urnes classique qu'est l'attachement préférentiel peut modéliser des propriétés universelles de réseaux complexes. En effet, cet artifice nous permet de suivre quels individus interagissent entre eux dans le cadre de leurs activités. Il est alors posside de faire ressortir certaines structures universelles dans les toiles d'interaction obtenues. Plus précisément, nous unifions ici l'*indépendance d'échelle*, la *modularité* et l'*auto-similarité* des réseaux complexes sous le concept d'organisation communautaire libre d'échelle.

Pour ce faire, nous offrons une nouvelle perspective d'organisation des réseaux complexes où les communautés, plutôt que les liens, jouent le rôle d'unités fondamentales. Nous montrons comment notre modèle simple peut reproduire certaines caractéristiques des réseaux sociaux et des réseaux d'information en prédisant leur structure communautaire. De façon plus importante, nous montrons comment leurs nœuds et leurs communautés sont interconnectés. L'auto-similarité de ces systèmes se manifeste alors dans les distributions de connexions observées entre noeuds et entre communautés.

Summary

We here show how introducing *coloured balls* in the classic preferential attachment *balls-in-urns* process models the emergence of some universal properties of complex networks. This artifice allows us to track which individuals interact as part of their activities. We then highlight universal features of the obtained web of interactions. More precisely, we here unify the *scale independence*, *modularity* and *self-similarity* of complex networks under the concept of scale-free community structure.

This brings a new perspective on network organization where communities, instead of links, act as the fundamental building blocks. We show how our simple model can reproduce certain features social and information networks by predicting their community structure. More importantly, we also show how their nodes and their communities are interconnected. The self-similarity of these systems then manifests itself in the distributions of connections observed between nodes and between communities.

5.1 Structural Preferential Attachment: networks beyond the link

5.1.1 A universal matter

Reducing complex systems to their *simplest possible form* while retaining their important properties helps model their behaviour independently of their nature. Results obtained via these abstract models can then be transferred to other systems sharing a similar simplest form. Such groups of related systems are called *universality classes* and are the reason why some models apply just as well to the sizes of earthquakes or solar flares than to the sales number of books or music recordings [78]. That is, their statistical distributions can be reproduced by the same mechanism: *preferential attachment*. This mechanism has been of special interest to network science [14] because it models the emergence of power-law distributions for the number of links per node. This particular feature is one of the universal properties of network structure [12], alongside modularity [48] and self-similarity [94]. Previous studies have focused on those properties one at a time [12, 49, 94, 3, 52, 95], yet a unified point-of-view is still wanting. In this chapter, we present an overarching model of preferential attachment that unifies the universal properties of network organization under a single principle.

Preferential attachment is one of the most ubiquitous mechanisms describing how elements are distributed within complex systems. More precisely, it predicts the emergence of *scale-free* (power-law) distributions where the probability P_k of occurrence of an event of order k decreases as an inverse power of k (i.e., $P_k \propto k^{-\gamma}$ with $\gamma > 0$). It was initially introduced outside the realm of network science by Yule [102] as a mathematical model of evolution explaining the power-law distribution of biological genera by number of species. Independently, Gibrat [47] formulated a similar idea as a law governing the growth rate of incomes. Gibrat's law is the sole assumption behind preferential attachment: the growth rates of entities in a system are proportional to their size. Yet, preferential attachment is perhaps better described using Simon's general *balls-in-bins* process [92].

Simon's model was developed for the distribution of words by their frequency of occurrence in a prose sample [104]. The problem is the following: what is the probability $P_{k+1}(i+1)$ that the $(i+1)$-*th* word of a text is a word that has already appeared k times? By simply stating that $P_{k+1}(i+1) \propto k \cdot P_k(i)$, Simon obtained the desired distribution [Fig. 5.1a]. In this model, the nature of the system is hidden behind a simple logic: the "popularity" of an event is encoded in its number of past occurrences. More clearly, a word used twice is 2 times more likely to reappear next than a word used once. However, before its initial occurrence, a word has appeared exactly zero times, yet it has a certain probability p of appearing for the very first time. Simon's model thus produces systems whose distribution of elements falls as a power law of exponent $\gamma = (2-p)/(1-p)$.

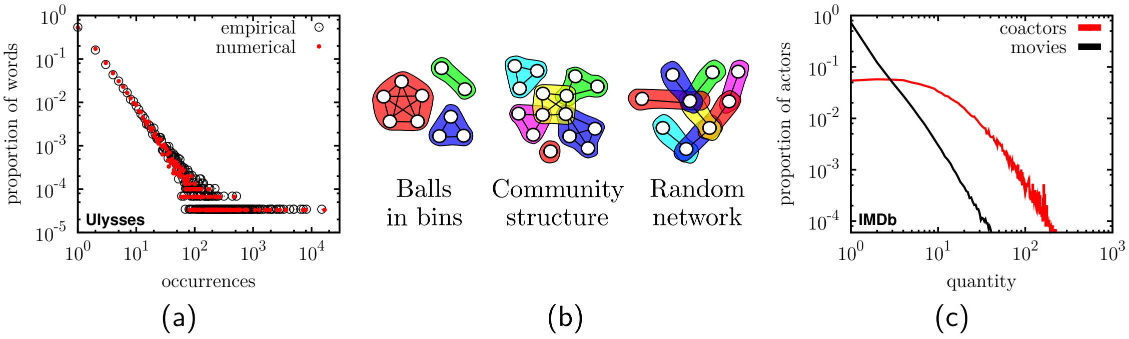


Figure 5.1 – **Spectrum of scale-free complex systems.** (a) The distribution of words by their number of appearances in James Joyce's Ulysses (empirical data). The numerical data was obtained from a single realization of Simon's model with p equal to the ratio of unique words (30 030) on the total word count (267 350). (b) Schematization of the systems considered in this chapter, illustrating how order (Simon's model of balls in bins) and randomness (Barabási-Albert's model of random networks) coexist in a spectrum of complex systems. (c) The distribution of co-actors and movies per actor in the Internet Movie Database since 2000. The organization moves closer to a true power law when looking at a higher structural level (i.e., movies versus co-actors).

5.1.2 On the matter of networks and community structure

Networks are ensembles of potentially linked elements called *nodes*. In the late 1990s, it was found that the distribution of links per node (the *degree distribution*) featured a power-law tail for networks of diverse nature. To model these so-called *scale-free networks*, Barabási and Albert [12] introduced preferential attachment in network science. In their model, nodes are added to the network and linked to a certain number of existing nodes. The probability that the new node chooses an old one of degree k is proportional to $k \cdot N_k$, where N_k is the number of nodes of degree k. As the system goes to infinity, N_k falls off as k^{-3}.

From the perspective of complex networks, Simon's model may be regarded not as a scheme of throwing balls (e.g., word occurrences) in bins (e.g., unique words), but as an extreme case of scale-free networks where all links are shared within clearly divided structures. Obviously, both Simon's and the Barabási-Albert's (BA) models follow the preferential attachment principle. However, Simon's model creates distinct growing structures, like the balls in bins of Fig. 5.1(a), whereas the BA model creates overlapping links of fixed size, as on the random network of Fig. 5.1(c). By using the same principle, one creates order while the other creates randomness [Fig.5.1b]. Our approach explores the systems that lie in between.

The vast majority of natural networks have a *modular topology* where links are shared within dense subunits [48]. These structures, or *communities*, can be identified as social groups, industrial sectors, protein complexes or even semantic fields [81]. They typically overlap with each other by sharing nodes and their number of neighbouring structures is called their *community degree*. This particular topology is often referred to as *community structure* [Fig. 5.1b]. Because these structures are so important on a global level, they must influence local growth. Consequently, they are at the core of our model.

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

Second, studies have hinted that Gibrat's law holds true for communities within social networks [91]. The power-law distribution of community sizes recently observed in many systems (e.g., protein interaction, word association and social networks [81] or metabolite and mobile phone networks [2]) supports this hypothesis.

Third, degree distributions can deviate significantly from true power laws, while higher structural levels might be better suited for preferential attachment models [Fig. 5.1c].

5.1.3 A simple generalization of Preferential Attachment

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

Our model can be described as the following stochastic process. At every time step, a node joins a structure. The node is a new one with probability q, or an old one chosen proportionally to its membership number with probability $1 - q$. Moreover, the structure is a new one of size s with probability p, or an old one chosen among existing structures proportionally to their size with probability $1 - p$. These two growth parameters are directly linked to two measurable properties: modularity (p) and connectedness (q) [Fig. 5.2]. Note that, at this point, no assumption is made on how nodes are linked within structures; our model focuses on the modular organization.

Whenever the structure is a new one, the remaining $s - 1$ elements involved in its creation are once again preferentially chosen among existing nodes. The basic structure size s is called the *system base* and refers to the smallest structural unit of the system. It is not a parameter of the model *per se*, but depends on the considered system. For instance, the BA model directly creates links, i.e. $s = 2$ (with $p = q = 1$), unlike Simon's model which uses $s = 1$ (with $q = 1$). All the results presented here use a node-based representation ($s = 1$), although they can equally well be reproduced via a link-based representation ($s = 2$). In fact, for sufficiently large systems, the distinction between the two versions seems mainly conceptual (see Ref. [57] for details).

In our process, the growth of structures is not necessarily dependent on the growth of the network (i.e., the creation of nodes). Consequently, we can reproduce statistical properties of real networks without having to consider the large-size limit of the process. This allows our model to naturally include finite size effects (e.g., a distribution cut-off) and increases freedom in the scaling properties. In fact, we can follow S_n and N_m, respectively, the number of structures of size n and of nodes with m memberships, by writing master equations for their time evolution [35]:

$$\dot{S}_n(t) = (1-p)\frac{(n-1)S_{n-1}(t) - nS_n(t)}{[1 + p(s-1)]\,t} + p\delta_{n,s}\;; \tag{5.1}$$

$$\dot{N}_m(t) = (1+p(s-1)-q)\frac{(m-1)N_{m-1}(t) - mN_m(t)}{[1 + p(s-1)]\,t} + q\delta_{m,1}\;. \tag{5.2}$$

Equations (5.1) and (5.2) can be transformed into ordinary differential equations for the evolution of the distribution of nodes per structure and structure per node by normalizing S_n and N_m by the total number of structures and nodes, pt and qt, respectively. One then obtains recursively the following solutions for the normalized distributions at statistical equilibrium,

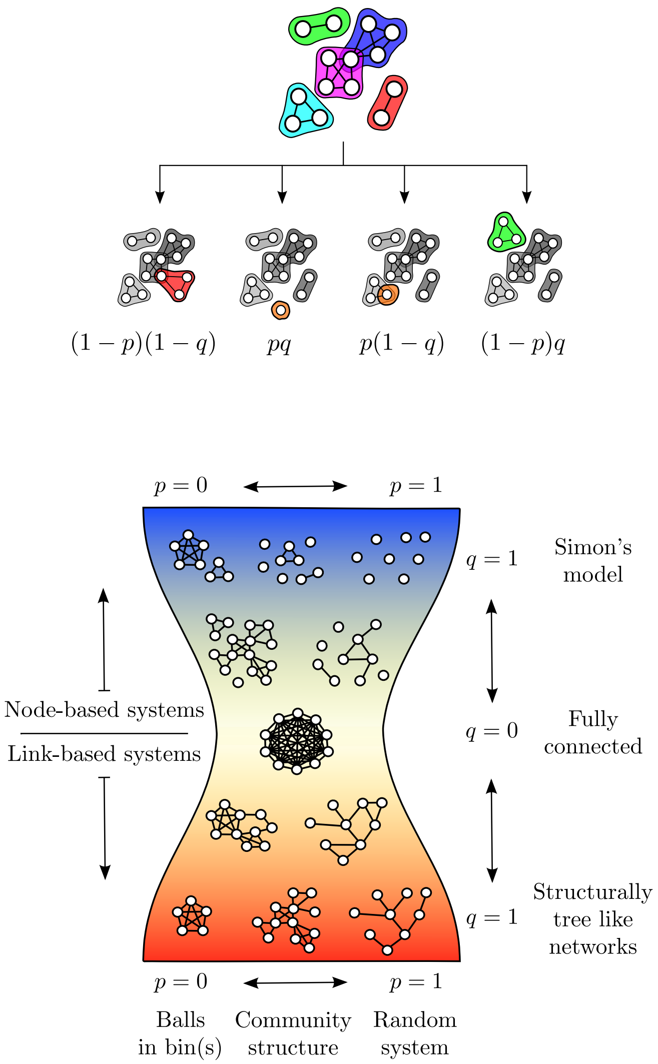


Figure 5.2 – **Structural preferential attachment and the systems it creates.** (top) Representation of the possible events in a step of node-based SPA; the probability of each event is indicated beneath it. (bottom) A schematization of the spectrum of systems obtainable with SPA. Here, we illustrate the conceptual differences between node-based $s = 1$ and link-based systems $s = 2$: Simon's model ($q = 1$) creates structures of size one (nodes), while the BA model ($p = q = 1$) creates random networks through structures of size two (links).

$\{\mathcal{S}_n^*\}$ and $\{\mathcal{N}_m^*\}$:

$$\mathcal{S}_n^* = \frac{\prod_{k=s}^{n-1} k\Omega_s}{\prod_{k=s}^{n}(1+k\Omega_s)} \quad \text{where} \qquad \Omega_s = \frac{1-p}{1+p(s-1)} \tag{5.3}$$

$$\mathcal{N}_m^* = \frac{\prod_{k=1}^{m-1} k\Gamma_s}{\prod_{k=1}^{m}(1+k\Gamma_s)} \quad \text{where} \qquad \Gamma_s = \frac{1+p(s-1)-q}{1+p(s-1)}\ , \tag{5.4}$$

which scale as indicated in Table 5.1, $\mathcal{N}_m^* \propto m^{-\gamma_N}$ and $\mathcal{S}_n^* \propto n^{-\gamma_S}$.

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

Table 5.1 – **Scaling exponents of SPA.** Exponents of the power-law distributions of structures per element (membership) and of elements per structure (size) at statistical equilibrium. One easily verifies that the membership scaling of link-based systems with $p=q=1$ corresponds to that of the BA model ($\gamma_N=3$), and that node-based systems with $q=1$ reproduce Simon's model.

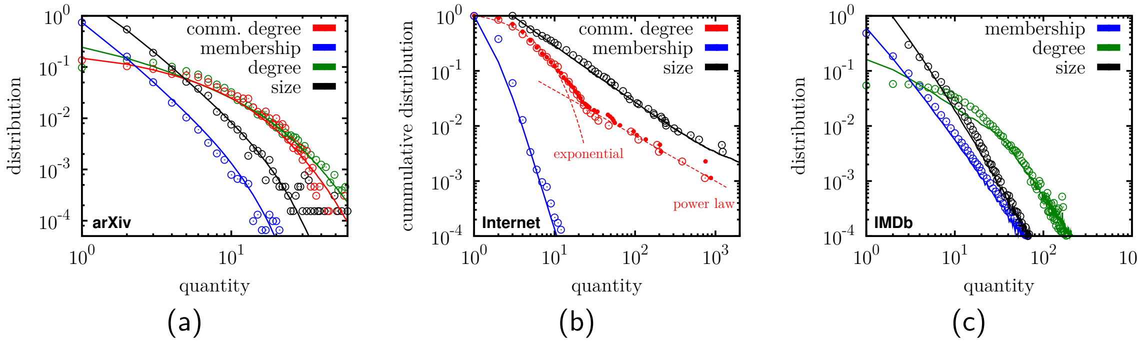


Figure 5.3 – **Reproduction of real systems with SPA.** Circles: distributions of topological quantities for (a) the *cond-mat arXiv* circa 2005; (b) Internet at the level of autonomous systems circa 2007; (c) the IMDb network for movies released since 2000. Solid lines: average over multiple realizations of the SPA process with (a) $p=0.56$ and $q=0.59$; (b) $p=0.04$ and $q=0.66$; (c) $p=0.47$ and $q=0.25$. For each realization, iterations are pursued until an equivalent system size is obtained. The Internet data highlights the transition between exponential and scale-free regimes in a typical community degree distribution. It is represented by a single realization of SPA (dots), because averaging masks the transition.

5.1.4 Results and discussions

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

The first set of results considered is the community structure of the co-authorship network of an electronic preprints archive, the *cond-mat arXiv* circa 2005 [Fig. 5.3a], whose topology was already characterized using a clique percolation method [81]. Here, the communities are detected using the link community algorithm of Ahn *et al.* [2], confirming previous results.

Using only two parameters, our model can create a system of similar size with an equivalent topology according to the four distributions considered (community sizes, memberships, community degree and node degree). Not only does SPA reproduce the correct density of structures of size 2, 3, 4 or more, but it also correctly predicts *how* these structures are interconnected via their overlap, i.e., the community degree. This is achieved without imposing any constraints whatsoever for this property. The first portion of the community degree distribution is approximately exponential; a behaviour which can be observed in other systems, such as the Internet [Fig. 5.3b] and both a protein interaction and a word-association network [81]. To our knowledge, SPA is the first growth process to reproduce such community structured systems.

Moreover, assuming fully connected structures, SPA correctly produces a similar behaviour in the degree distribution of the nodes. Obtaining this distribution alone previously required two parameters and additional assumptions [3]. In contrast, SPA shows that this is a signature of a *scale-free community structure*. This is an interesting result in itself, since most observed degree distributions follow a power law only asymptotically. Furthermore, this particular result also illustrates how self-similarity between different structural levels (i.e., node degree and community degree distributions) can emerge from the scale-free organization of communities.

Finally, the Internet Movie Database co-acting network is used to illustrate how, for bigger and sparser communities which cannot be considered fully connected, one can still easily approximate the degree distribution. We first observe that the mean density of links in communities of size n approximately behaves as $\log(n)/n$ (see Sec. 5.2). Then, using a simple binomial approximation to connect the nodes within communities, it is possible to approximate the correct scaling behaviour for the degree distribution [Fig. 5.3c]. This method takes advantage of the fact that communities are, by definition, homogeneous such that their internal organization can be considered random.

5.1.5 Conclusion and perspective

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

Looking at network organization beyond the link is also useful to account for missing links [27] or to help realistic modelling [62, 64]. For instance, this new paradigm of scale-free community structure suggests that nodes with the most memberships, i.e., structural hubs, are key elements in propagating epidemics on social networks or viruses on the Internet. These structural hubs connect many different neighbourhoods, unlike standard hubs whose links can be redundant if shared within a single community. Consequently, local action on these nodes can be a more globally effective way to control a given dynamics [58].

There is no denying that communities can interact in more complex ways through time [80]. Yet, from a statistical point-of-view, those processes can be neglected in the context of a structurally preferential growth. Similarly, even though other theories generating scale-free designs exist [37], they could also benefit from generalizing their point of view to higher levels of organization.

5.2 Appendix 5.A: Data, details and methods

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

5.2.1 Data

Internet Movie Database The dataset used for the co-acting network of IMDb consists only of movies released after December 31st 1999. Interestingly, the degree distribution is almost identical to that published a decade earlier [12] which consisted of all movies released before the turn of the century. This suggests, since the two networks contain distinct and exclusive ensembles of movies, that the growth parameters of the IMDb network are constant. The network contains 7 665 259 links between 716 463 nodes (actors), where two actors share a link if they are credited alongside another for at least one movie. It was only analysed using the link community algorithm, because of memory issues with CFinder. The organization levels corresponding to actual movies, which is how the dataset was originally compiled, was deemed unsuitable for the study because of the presence of economic (limiting the number of actors in a movie) and artistic (typically requiring a minimal number of characters in a movie) constraints. We believe that a community detection process on the network actually frees the system from these constraints and yield communities of actors linked by genre, time, location, etc.

arXiv The cond-mat arXiv database uses articles published at *http://arxiv.org/archive/cond-mat* between April 1998 and February 2004. In this network, an article written by n co-authors contributes to a link of weight $(n-1)$ between every pair of authors. The unweighted network was obtained by deleting all links with a weight under the selected threshold of 0.1; resulting in a network of 125 959 links between 30 561 nodes (authors). This dataset was compiled, analysed and presented in [81].

Internet This dataset is a symmetrized snapshot of the structure of the Internet at the level of autonomous systems, reconstructed from BGP tables posted at *archive.routeviews.org*. This snapshot was created by Mark Newman from data for July 22nd 2006. The network contains 22 962 nodes and 48 436 links.

5.2.2 Community detection

Community detection in networks is a challenge in itself. For instance, SPA creates network when communities overlap. Consequently, we require algorithms that can 1) assign nodes to more than one communities and 2) be tuned to investigate different levels of organization to find one suited to be modelled through preferential attachment. In order to characterize the networks used in Sec. 3.1, two independent and completely different algorithms were used: a link community algorithm [2] and the clique percolation method of CFinder [81]. Results use the link community algorithm, because it proved to be faster and better suited to detect communities within communities. When possible, CFinder was used for cross-checking the community partition.

Link communities

This algorithm assigns links, instead of nodes, to communities. Two links, e_{ij} and e_{ik}, stemming from a given node i are said to belong to the same community if their Jaccard coefficient $J(e_{ij}, e_{ik})$ (similarity measure) is above a given threshold J_c :

$$J_{A,B} = \frac{A \cap B}{A \cup B} > J_c \ , \tag{5.5}$$

where A (B) is the set containing the neighbours of j (k) including j (k).

A large community can thus be composed of different smaller communities where the similarity of their members' neighbourhoods is higher than in the larger community. The link community algorithm proved to be quite efficient at detecting these nested communities. That being said, we have since developed a naive but simple and efficient way to get even better results from the same algorithm [101].

CFinder and clique percolation

The original clique percolation method used by CFinder is designed to locate the k-clique communities of unweighted, undirected networks. This community definition is based on the observation that a typical member in a community is linked to many other members, but not necessarily to all other nodes in the community. In other words, a community can be interpreted as a union of smaller complete (fully connected) sub-graphs that share nodes.

Such complete sub-graphs in a network are called k-cliques, where k refers to the number of nodes in the sub-graph, and a k-clique-community is defined as the union of all k-cliques that can be reached from each other through a series of adjacent k-cliques. Two k-cliques are said to be adjacent if they share $k-1$ nodes. CFinder is available at `http://cfinder.org/`.

5.2.3 Levels of organization

The first step when looking to compare the structure of real networks with systems produced by SPA is to analyse the empirical data. As mentioned earlier, our main algorithm (the link community algorithm [2]) has a single parameter to tune for community detection: its Jaccard threshold. The Jaccard threshold embodies *how similar* the neighbourhoods of the ends of two links must be in order for these links to be considered as part of the same link community. Tuning this parameter, demanding how tightly connected a group of nodes must be in order to be labelled as a community, allows us to look at different levels of organization within the network. If too small, the algorithm will most likely end up with communities corresponding to the connected components of the networks. If too big, significant communities will be broken up into different smaller ones. In this work, we proceeded by sweeping this parameter in order to find the level of scale-free organization.

5.2.4 A note on node-based and link-based systems

All results presented in the SPA section used a node-based version of SPA. Which means that new structures contain a single node and that they will remain disconnected from the other components of the network until they reach an older node. For the IMDb data, this choice is not even a question as the network contains many such satellites structures (even some of size one) which are disconnected from the giant component. In other systems, like the arXiv network, the choice can be more complicated. One might be tempted to use a link-based system process to reproduce the arXiv, since it is a co-author network and thus cannot contain isolated nodes. However, it does contain some disconnected components, which a link-based process like the Barabási-Albert model [12] is incapable of producing. Hence, it seemed logical to use the node-based process and simply remove the structures of size one (nodes who failed to co-author a paper) from the final system. As a final point on the subject, it is interesting to note that we have been able to reproduce all results using both the node-based and link-based version of SPA. In sufficiently large and connected systems, the distinction between the two seems mainly conceptual. In fact, a naive mapping of one to the other has been shown to be very accurate [57].

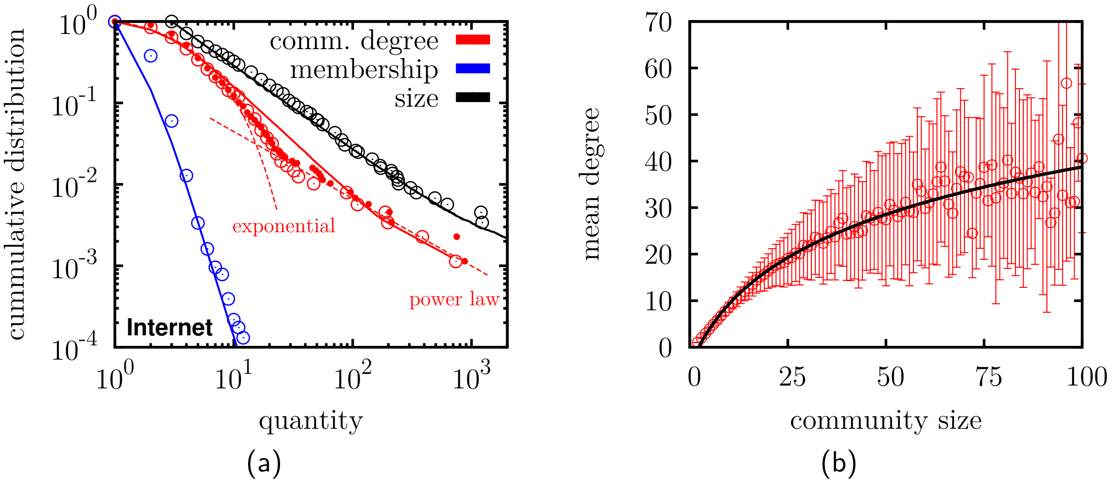


Figure 5.4 – **Supplementary results for SPA on the Internet and the IMDb.** (a) ⊙: distributions of topological quantities for the ensemble of the Internet at the level of autonomous system circa 2007; solid lines: average over multiple realizations of the SPA process with $p = 0.04$ and $q = 0.66$. The empirical network was analysed using the link community algorithm [2] with Jaccard threshold 0.08. (b) The mean number of links per node within a given community as a function of the community size in the IMDb network. The fit is done using a logarithmic function of the form $f(x) = a \cdot log(x+b) - c$.

5.2.5 Supplementary results and discussions

Section 3.1 presented our results for the arXiv network, the Internet and the Internet Movie Database. The arXiv data is completely shown, but the Internet is illustrated for communities of size 3 or bigger (as done by the authors of the detection algorithm [2]) because the algorithm can overestimate the number of communities of size 2 and the goal is here to highlight the connectedness of communities. For the IMDb, the community size distribution is normalized for communities of size 3 or bigger, but the communities of size 2 are considered in the membership and degree distributions. These results highlight how these systems follow a scale-free community structure and how SPA can be used to predict behaviour *outside* of the model's specification. More precisely, the numerical systems predict how the communities are interconnected via their overlap, reproducing the exponential behaviour and the heavy tail of the community degree distribution. It is interesting to note that averaging over many iterations of the SPA process highlights the distribution cut-off caused by the finite size of the system. This effect is mostly visible on the arXiv results. On the other hand, because the position of the transition between exponential and power-law behaviour observed in the cumulative community degree distribution is highly dependent on the amount of "leading" structures (i.e. the number of structures which are able to break away from the majority and thus have a significantly bigger size), it can differ slightly between two realizations of SPA. In this context, averaging over multiple iterations of the same process partly smooths out the transition. For this reason, a single realization of the model is also presented on Fig. 5.4a to better illustrate the behaviour of community degree in a finite system.

5.2.6 From communities, back to links

This last subsection presents results which, although preliminary, imply that individuals within a given social community can be approximated as being randomly connected. The first step in shifting our point of view from communities back to links is to evaluate just how connected the communities of our systems are. Figure 5.4b illustrates the mean number of links per node within a given community as a function of the community size, which is found to grow logarithmically. Using this measure to determine the density of a structure of a given size, we simply throw a dice for each possible link to determine which links actually exist, while respecting the actual density of the network. This allows us to move from a potential degree distribution to an estimated degree distribution. If the binomial approximation (all links within a given community exist only with a certain probability) is correct, this estimated degree distribution should be close to the actual degree distribution of the system we are trying to reproduce. According to Fig. 5.3c, this is indeed the case. It is easy to note that the number of nodes of small degree is overestimated by SPA. As we will see in the next chapter, this is most likely a result of not considering enough structural levels.

5.3 Appendix 5.B: A preview of a future application

This chapter dealt with the so-called community structure of networks: a description of the grouping of nodes in communities in terms of group size and overlap. However, except for very basic assumptions, we never attempted to describe the *structure of communities*: an internal description of communities as sub-networks. In fact, the community structure literature is far from reaching any consensus on how to define these communities. How dense or clustered are they? How flexible must our definition be for it to remain applicable in a scale-independent community structure? In one of our ongoing projects, we attempt to answer these questions in order to generalize SPA to a model of community structure *and* structure of communities. To do so, we focus mainly on empirical observations.

5.3.1 Empirical observations

As mentioned in introduction, even though the field of community detection on networks is blooming, there are currently no consensus on what a community should look like. Consequentially there is no clear definition of what models of network growth should aim for when modelling the structure of communities. That being said, some common features can be extracted from almost all definition of communities, such that one could hope to design a

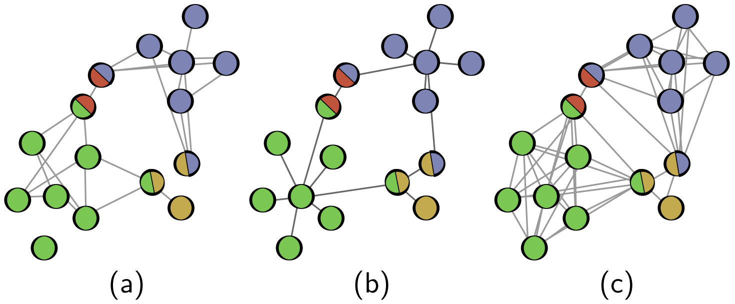


Figure 5.5 – **Examples of unacceptable structure of communities.** Although unacceptable on their own, our mechanism implies a flexible definition encompassing the variable density of (a), the potential heterogeneity of (b) and the connectedness of (c).

model flexible enough to be useful in association with most detection algorithms. Figure 5.5 presents a few possible organization of communities. Yet, all of these organisations can and should be rejected based on intuitive properties of social groups.

First, Fig. 5.5a presents communities formed through a Poissonian scheme (or Bernouilli trial) as in Erdős-Rényi graphs [41], where each pair of nodes are potentially linked with a fixed probability. This scheme does not reject the possibility of having nodes of degree zero. Intuitively, a community with individuals sharing no links with the other members does not make much sense. Similarly, with the network as sole information, no algorithm could potentially assign a node to a group in which he share no connection. Our first consideration thus implies community connectedness: each member of a community should be reachable from any other member.

Second, Fig. 5.5b presents highly heterogeneous communities. Community structure, under its simplest form, is defined in opposition to random organization; implying correlations between one's friends, and the friends of his friends. Community can be heterogeneous, but they must at least be distinguishable from the rest of the network. This second consideration requires a certain redundancy between the neighbourhoods of members of a given community, and consequently, that small community should be denser than expected in a corresponding fully rewired (or random) network.

Third, Fig. 5.5c presents fully connected communities. While these communities respect our first two considerations. They are obviously far too rigid to be of practical use. For instance, in a network of professional ties where we would seek to identify different fields and companies, we would never require a given individual to be connected with every single other employee of his company. Obviously, while a complete density might be expectable for smaller groups, it is hardly ever achieved in larger social structure (companies, universities, cities). This third consideration merely relaxes the second one: large communities can be sparse even while featuring redundancy in the neighbourhoods of their members.

Based on these three considerations and the SPA model, we can go forward and propose a simple and intuitive mechanism describing how social groups tend to grow.

5.3.2 Including the structure of communities in SPA

A preferential attachment scheme at the level of communities implies that a community has a growth rate (or recruitment rate) proportional to its current size. This can be interpreted as if each current member of a community introduces a new member at a given rate. This preferential attachment mechanism is not only intuitive, but does in fact reproduce properties of real networks [56]. It is then natural to consider a similar logic for the creation of new links between members of a community.

Our generalized SPA* model is simply stated. The growth of communities is governed by our SPA model, but we include an additional process for the growth of structure *within* communities. From an internal point of view, each member of a community of size n recruits a new member at rate r, such that the growth rate $\dot{n}$ of a community of size n is proportional to rn. The new member is initially connected only to the individual who introduced it within the group (its degree, i.e. number of links, equals 1 within this community). This assures connectedness (consideration #1). However, each member also creates a new link at a rate b until its degree equals $n-1$ (such that it is connected to every other member). The links created by existing members must then randomly select a receiving individual. Consequently, a single member can gain degree faster than rate b if other members make the effort of creating the link, thus helping smaller communities maintain a high density (consideration #2).

The number $n_k(t)$ of individuals with degree k within an average community of size n can be followed through time t by a master equation:

$$\begin{aligned}\dot{n}_k(t) &= rn\delta_{k,1} + b\left(n_{k-1}\bar{\delta}_{k,n} - n_k\bar{\delta}_{k,n-1}\right) + r\left(n_{k-1} - n_k\right) \\ &\quad + \left(b\sum_{k'=1}^{n-2} n_{k'}\right)\frac{(n-k)\,n_{k-1} - (n-1-k)\,n_k}{\sum_{k'=1}^{n-1}(n-1-k')\,n_{k'}}\end{aligned} \tag{5.6}$$

where $\delta_{i,j}$ is the Kronecker delta and $\bar{\delta}_{i,j} = 1 - \delta_{i,j}$. The first term accounts for the arrival of new members (with $k = 1$) at rate rn. The second term is due to the creation of links, which brings an individual of degree $k-1$ to degree k [positive for $n_k(t)$], and individual of degree k to degree $k+1$ [negative for $n_k(t)$]. The third term is due to the receiving end of links created when a new individual joins the community; while the last term accounts for the link creation between existing members. The parenthesis is the creation rate, while the ratio yields the probabilities of it affecting a node of degree $k-1$ [positive for $n_k(t)$] or of degree k [negative for $n_k(t)$]. This complete description of the average state of a community is validated in Fig. 5.6.

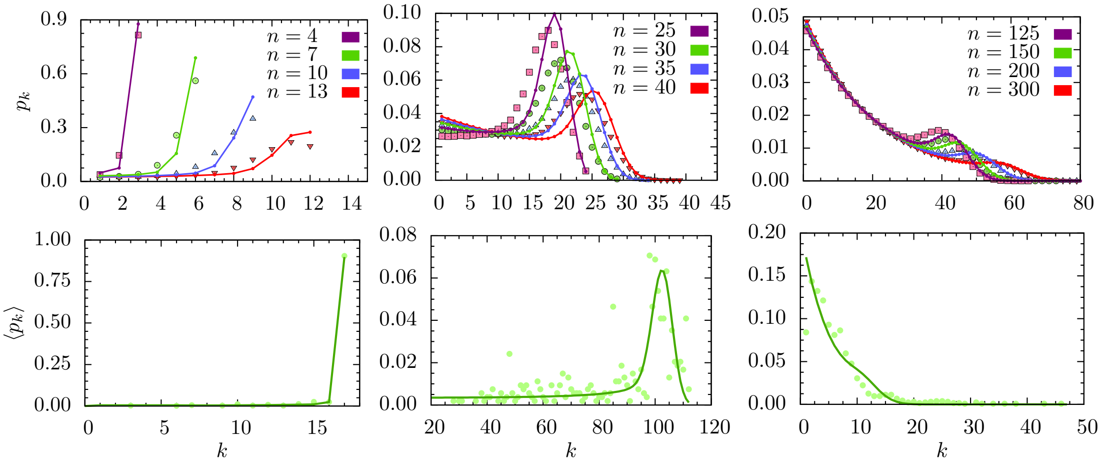


Figure 5.6 – **Internal structure of diverse communities in SPA* and empirical data.** (top) Degree distribution for various community sizes with $\lambda = 9$. Comparison between results of Eq. 5.6 (small dots) and the average of Monte-Carlo simulations (closed squares, circles and triangles). Lines are added to guide the eye. Small and medium size communities are highly homogeneous, while degree distributions of larger communities are heavily skewed: with a mode at $k = 1$ for an average approaching $\langle k \rangle_n = 20$ in the limit $n \to \infty$. The discrepancy between simulations and the results of Eq. 5.6 can be traced back to the continuous approximation involved in writing differential mean-field models, as well as the absence of structural correlation in this type of model. The net effect is a shift of the prediction toward higher degrees for the bulk of the distribution. (bottom) Similar organization found in real-world networks: (left and middle) small and medium size communities found in the sexual network presented in Chap. 3, (right) large scale communities found in the co-authorship network presented in this chapter. Lines again follow the mean-field description of the model, with λ chosen by hand to qualitatively reproduce the correct internal state. The empirical distributions are obtained by averaging over multiple communities of similar size.

A simpler point of view can be adopted to gain further insights into the relation between the internal degree of an average individual (node) $\langle k \rangle$ and the size $n(t) = \sum_{k'} n_{k'}(\mathrm{t})$ of the community. Since the average internal degree $\langle k \rangle_n$ is directly related to the average number of links $L(t)$ within a community of size $n(t)$ at time t, we obtain a mean-field equation for the latter as it is easier to follow analytically. Assuming a uniform and uncorrelated distribution of links among individuals, we can write

$$\frac{dL}{dt} = rn + bn\left[1 - \left(\frac{L}{L_{\max}(n)}\right)^{n-1}\right] , \tag{5.7}$$

since any given individual will create a link at rate b *if* he has currently less than $n-1$ links. Here $L_{\max}(n)$ simply equals $n(n-1)/2$ (the case where every link exists). A straightforward transformation yields L as a function of the average size at time t (using $dn/dt = rn$):

$$\frac{dL}{dn} = \frac{dL}{dt}\frac{dt}{dn} = 1 + \lambda\left[1 - \left(\frac{L}{L_{\max}(n)}\right)^{n-1}\right] , \tag{5.8}$$

where $\lambda = b/r$. While the degree distribution is neither uniform nor uncorrelated in the complete model (see Fig. 5.6), we will see that our simplification are robust enough, and that Eq. (5.8) accurately reproduces the average density of communities.

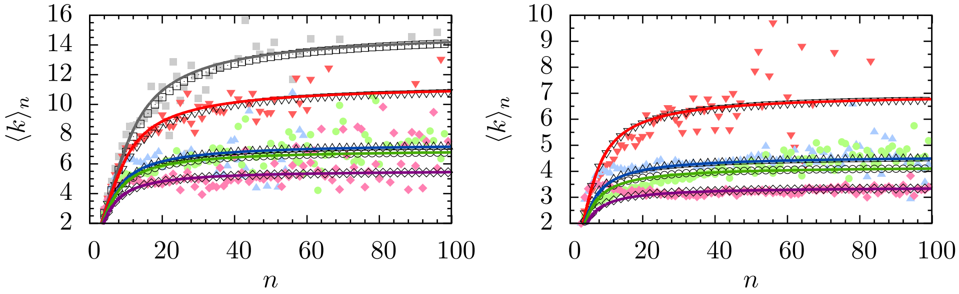


Figure 5.7 – **Relation between SPA* and "Dunbar's number".** We show the average number of connections for a given individual within a group of size n. (top) arXiv (bottom) MathSciNet [83]. The community detection algorithms are: (gray squares) cascading approach to clique percolation (CCPA) [101, 81], (red inverted triangles) link community algorithm (LCA) [2], (blue triangles) order statistics local optimization method (OSLOM) [68], (green circles) greedy modularity optimization on line-graphs (LG) [45], and (magenta lozenge) greedy clique expansion (GCE) [69]. All algorithms are used as prescribed by their original authors: CCPA is performed with k-cliques of sizes 6,4, and 3; the maximal density partition that excludes singleton is selected for LCA, the basic level of organization (community structure) is selected for OLSOM, the random-walker mapping with weights and self loops is used for LG, and we use default parameters for GCE (4 is the minimal clique size). Analytical predictions of Eq. 5.8 are shown in solid lines, tuned for the organization represented by each algorithm, with averages of the corresponding Monte-Carlo simulations in colorless symbols. All values of λ were set using Eq. 5.10 and are, from top to bottom: (arXiv) 6.5, 4.7, 2.7, 2.5 and 1.8. (MathSci) 2.5, 1.3, 1.1 and 0.7.

A simple analysis of Eq. (5.8) highlights an interesting feature of this model. For large sizes n, the ratio $L/L_{\max(n)}$ goes to zero while $n \gg 1$ by definition, such that a maximal link creation rate

$$\frac{dL}{dn} \simeq 1 + \lambda \tag{5.9}$$

is attained. Hence, the intensive quantity $L/n \to (1 + \lambda)$ converges toward a constant that depends on the parametrization of the model alone. Considering that one link equals two stubs (or degree), the asymptotic average degree is directly related to the parameter λ through:

$$\langle k \rangle_\infty = \frac{2L}{n} = 2\,(1 + \lambda) \ . \tag{5.10}$$

This indicates a maximal average number of connections in social group (consideration #3).

5.3.3 Relation to Dunbar's number

Previous results and those presented in Fig. 5.7 illustrate how different community detection algorithms possess qualitatively equivalent features even if they use different definitions of communities and consequently investigate different organizations. More precisely, results of Fig. 5.7 highlight how there exist two different behaviours for the average number of links per individual in relation to the size of a social group. For low sizes n, the mean degree $\langle k \rangle_n$ essentially scales linearly with the community size as *everybody knows each other* *within small groups* (e.g. family or close friends). For larger sizes n, $\langle k \rangle_n$ reaches a plateau [1], Eq. (5.10), where a typical individual will not gain new connections when the potential number of connections is increased. So while there is no maximal community size per se, there is a *maximal number of connections that an average individual might possess within a given group* (e.g. large companies or online communities). This effectively implies a maximal community size as connection density will decrease as size is increased until the community barely qualifies as a group.

Interestingly, this behaviour of individual activity $\langle k \rangle_n$ with respect to community size n has been previously observed in studies related to an anthropological theory known as Dunbar's number [40]. This theory is based on the observed relation between neocortical sizes in primates and the characteristic sizes of their social groups. Its interpretation usually involves information constraints related to the quality of interpersonal connections and our ability to maintain such relationships. While the importance of neocortical sizes is surely disputable [31], the fact remains that empirical evidence supports the existence of an upper bound in the absolute number of active relationships for a given individual [51]. Similarly, our empirical results indicate that an upper bound may exist in the number of connections one individual can maintain within a given social group.

In our intuitive model, this upper bound naturally emerges and is solely dependent on the λ parameter. This parameter can be interpreted as the ratio between the involvement of an individual in a community, in the sense of bonding with other members, and its contribution to the growth rate of the community. For low λ or large community sizes, the rate of change in the population is higher than an individual's involvement, such that the maximal degree stagnates. Whereas, for high λ and small communities, the individual is able to follow population changes and hence create relationships with most of its members. Thus, different types of social groups will feature different λ and consequently different values of "Dunbar's number".

In this interpretation, the upper bound for individual degree in social groups is due to the fact that connections and introduction of new members have linear requirements for individuals, whereas this implies that groups grow exponentially. Other mathematical models exist to describe Dunbar's number (e.g. [51]), usually based on arguments of priority and/or time resources. However, our model is based on the *observed* community structure of real world networks and consequently, explains Dunbar's number in terms of its two base units, individuals and groups, and the ratio of their respective characteristic rates.

Future work will investigate the use of this complete model as a benchmark for community detection algorithm. Investigating, for instance, how Dunbar's number implies a limit in detectability for communities of very large sizes as they become sparser than the global density of the network.

[1] The transition between the two density regimes was initially observed in Fig. 5.4b.

Chapter 6

On structure II: Hierarchical preferential attachment

Résumé

Les systèmes complexes réels n'ont pas de structure fixe, et il n'existe aucune règle pour leur reconstruction. Tout de même, dans leur apparent désordre, d'étonnantes propriétés structurelles semblent émerger universellement. Dans ce dernier chapitre, nous proposons qu'une large classe de systèmes complexes peut être modélisée comme une construction de plusieurs, voire une infinité, de niveaux d'organisation suivant tous le même principe universel de croissance qu'est l'attachement préférentiel. Nous donnons des exemples de telles hiérarchies, dont la pyramide de production de l'industrie cinématographique. Nous démontrons aussi comment les réseaux complexes peuvent être interprétés comme une projection de notre modèle, de laquelle leur indépendance d'échelle, leur modularité, leur hiérarchie, leur fractalité et leur navigabilité émergent naturellement. Nos résultats suggèrent que certaines propriétés des réseaux complexes peuvent être relativement simples, et que leur complexité apparente est largement une réflexion de la structure hiérarchique complexe de notre monde.

Summary

Real complex systems are not rigidly structured; no clear rules or blueprints exist for their construction. Yet, amidst their apparent randomness, complex structural properties appear to universally emerge. In this last chapter, we propose that an important class of complex systems can be modelled as a construction of potentially infinitely many levels of organization all following the same universal growth principle known as preferential attachment. We give examples of such hierarchy in real systems, for instance in the pyramid of production entities of the movie industry. More importantly, we show how real complex networks can be interpreted as a projection of our model, from which their scale independence, their clustering, their hierarchy, their fractality, and their navigability naturally emerge. Our results suggest that certain properties of complex networks can be quite simple, and that the apparent complexity of their structure is largely a reflection of the hierarchical nature of our world.

6.1 What makes complex networks complex

The science of complexity is concerned with systems displaying emerging properties; i.e., systems where the properties of the parts do not directly dictate the properties of the whole [93]. In what follows, we generalize the work of the previous chapter and show how one property of the whole, *hierarchy*, can alone be the origin of more complex features. We describe hierarchical systems through a general model of *coloured balls in embedded bins* which itself explains the emergence of other features through the projection of these hierarchical systems onto complex networks.

Real networks tend to be sparse and appear highly disorganized, but they also tend to feature properties not found in any classic models of sparse random networks: *scale independence*, fat-tailed degree distribution [24, 12]; *modularity*, the grouping of nodes in denser groups [98, 49, 56]; *hierarchy*, the embedding of multiple levels of organization [90, 27]; *fractality*, the self-similarity between levels of organization [94, 95]; and *navigability*, the possibility of efficient communication through a hidden metric space [20, 21, 84].

Sophisticated algorithms can be designed to reproduce most of these features, often based upon a multiplicative process to force their emergence by reproducing a basic unit on multiple scales of organization [89, 82]. These models are useful as they can create realistic structures and test hypotheses about measured data. However, these constructions are not intended to provide any insights on the underlying mechanisms behind a system's growth.

On the other hand, simple generative models are quite successful at suggesting principles of organization leading to specific properties. For example, simple models exist to propose possible origins for scale independence [12] or of the small-world effect [98], but they fail to model the emergence of properties not included by design. Consequently, the identification of new universal properties require the creation of new generative models, such that a single unifying principle has yet to be proposed.

In this last chapter, we bridge the gap between complex deterministic algorithms and simple stochastic growth models. To this end, we propose that to accurately model a complex network, we should first ignore it. The hierarchical nature of networks suggests that the observed links between nodes are merely projections of higher structural units [27, 56, 57] (e.g., people create groups within cities in given countries). These subsystems will be our focus. We use one general assumption to design an equally general model of hierarchical systems: embedded levels of organization all follow preferential attachment. We validate this model on the well documented dataset of production entities in the movie industry (i.e., producers produce films within companies in given countries). We then study the structure of the projection of this system onto a complex network of co-production between film producers. Interestingly, the resulting networks feature a scale-independent hierarchical organization, community structure, fractality and navigability.

By reducing the complex to the simple, we provide new insights on the mechanism behind the growth of complex networks at the level of individual nodes and open new perspectives on the origins of networks as we know them.

6.2 Hierarchical Preferential Attachment (HPA)

6.2.1 A return to classical preferential attachment

The preferential attachment principle is a ubiquitous *right-get-richer* mechanism modelling complex systems of all nature [102, 47, 92, 32, 24, 12, 56]. Simply put, it implies that the likelihood for a given entity to be involved in a new activity is roughly linearly proportional to its past activities. For instance, an individual with 10 acquaintances in a social network is roughly 10 times more likely to gain a new connection than one with a single acquaintance. This simple mechanism leads to a scale-independent organization of the distribution of the activity in question; modelling any system where the distribution of a resource among a population roughly follows a power-law distribution. Consequently, the number N_s of individuals with a share s ($\in \mathbb{N}$) of the resource scales as $s^{-\gamma}$, where γ is the scale exponent.

We consider a discrete time process where, during an arbitrary time step Δt, a new element i of share $s_i = 1$ is introduced within the system with probability B (birth event) and that the share s_j of one existing element j is increased to $s_j + 1$ with probability G (growth event). We can write a mean-field equation governing the number of individuals N_s with a given share s:

$$N_s(t + \Delta t) = N_s(t) + B\delta_{s,1} + \frac{G}{(B+G)\,t}\left[(s-1)\,N_{s-1}(t) - sN_s(t)\right] \tag{6.1}$$

where $(B+G)\,t = \sum sN_s(t)$, the sum of all shares, is used to normalize the transition probabilities. This simple model is easily shown to converge toward the following asymptotic organization

$$\lim_{t,s\rightarrow\infty} N_s(t) \propto s^{-\gamma} \quad \text{with } \gamma = 2 + \frac{B}{G}\,. \tag{6.2}$$

Considering that this organization is found in distributions of friends [12], of members in social groups [56] and of city population [104], it is natural to ask: How would a preferential attachment occurring on multiple levels influence the created structure? It is a popular idea that complexity frequently takes the form of hierarchy and that a hierarchical organization influences the property of the whole independently of the nature of its content [93]. With the recent successes of preferential attachment models, we hereby propose a generalization for hierarchical systems.

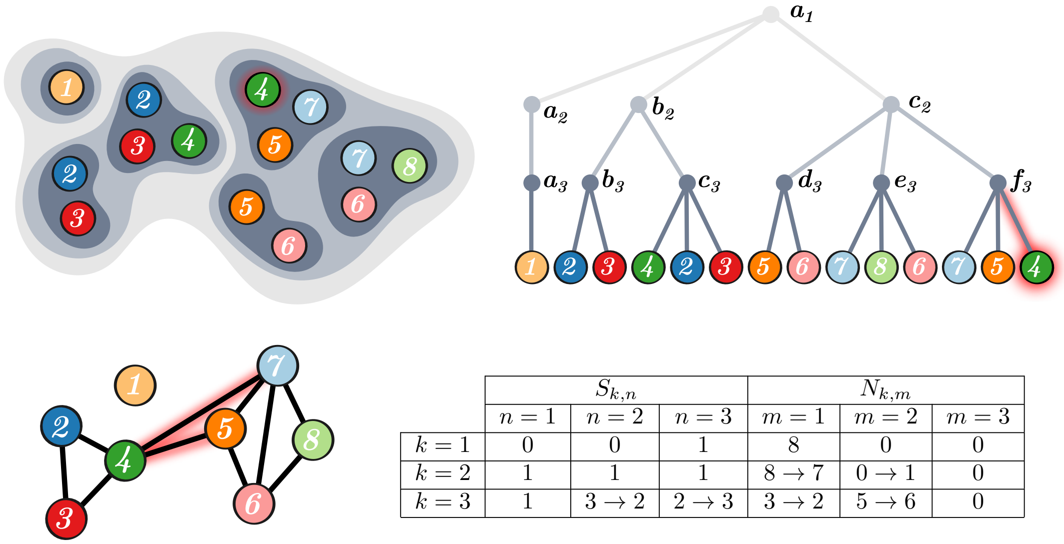
6.2.2 Hierarchical preferential attachment (HPA)

Classic preferential attachment processes can be described as schemes of throwing balls in bins. In the last chapter, we introduced coloured balls to represent individuals in social systems where individuals (unique colours) can be interpreted as a resource of social groups (boxes) and vice versa [56]. This leads to a natural interpretation of preferential attachment as a growth process for structured systems. Here, we generalize further by considering systems consisting of an arbitrary number d of embedded levels of organization. Hence, we can describe HPA as a scheme of throwing coloured balls in d embedded levels of structures (say, urns in bins in boxes for $d = 3$).

HPA processes can be described by using d different versions of Eq. (6.1) for the sizes of structures (e.g., how many balls in each urn? or how many urns in each bin?) and d more for the activities of individuals (e.g., in how many urns/bins/boxes does a given colour appear?). The dynamics is then completely determined, assuming we obtain the birth and growth probabilities (i.e., B and G for colours and structures at each structural level).

6.2.3 A model of HPA

We here describe a simple HPA model based on Herbert Simon's preferential attachment process [92] and explicitly show how it can be followed analytically. However, note that the details involved in the analytical description of the model are not necessary to our results.



	$S_{k,n}$			$N_{k,m}$		
	$n=1$	$n=2$	$n=3$	$m=1$	$m=2$	$m=3$
$k=1$	0	0	1	8	0	0
$k=2$	1	1	1	$8\rightarrow 7$	$0\rightarrow 1$	0
$k=3$	1	$3\rightarrow 2$	$2\rightarrow 3$	$3\rightarrow 2$	$5\rightarrow 6$	0

Figure 6.1 – **Schematization of hierarchical preferential attachment.** HPA process frozen as a green ball (labelled 4) is added to structure f_3. In this event, structure a_1 was chosen for growth (probability $1-p_1$), then structure c_2 (probability $(1-p_2)\cdot 3/6$), then structure f_3 (probability $(1-p_3)\cdot 3/8$). The colour had to be new for structure f_3 ($q_3=1$), and was chosen to be new for structure c_2 (probability q_2), but old for structure a_1 (probability $1-q_1$). At this point, the accessible colours were those labelled 1, 2, 3 and 4 and all had the same probability of being chosen as they all belong to a single level 2 structure. (*top left*) Representation as coloured balls in embedded levels of structure (urns in bins in boxes). (*top right*) Hierarchical representation as an inverted tree. Navigating downward corresponds to moving in smaller structure until we reach the balls therein. (*bottom left*) Network representation of links between balls. In this case, two nodes share an edge if they belong to a same level 3 structure. Other projection could be considered: for example, the network of level 2 structures sharing at least one type of ball. In this case, adding ball number 4 to structure f_3 would connect b_2 and c_2 while a_2 remains disconnected. (*bottom right*) Mathematical representation according to the quantities defined in the text: numbers $N_{k,m}$ of balls appearing in m structures of level k and numbers $S_{k,n}$ of level k structures containing n substructures (or ball) at level $k+1$.

Some visual representations of the model are given in Fig. 6.1. While cryptic at this point, they should become clearer with the following description.

During a time step Δt, one event takes place: a ball is thrown in d embedded structures. We refer to level 1 as the top, or superior, level (i.e., the biggest boxes). Consequently, this event marks the birth of a new structure at level 1 with probability p_1, or the growth of an existing level 1 structure with complementary probability $1-p_1$. If a new structure is created, this forces the creation of one structure on all inferior levels. If an existing structure is chosen for growth, this is done preferentially to their size, i.e., the number of level 2 structures that they contain. Within the chosen structure, the process is repeated. A new level 2 structure is created with probability p_2 or an existing one grows with probability $1-p_2$. Once the level (the smallest structure, the urn) in which the event occurs has been chosen — which implies that either a level x structure has been created or that we have reached the lowest level ($x=d$) — the colour of the involved ball must then be determined.

With probability $1 - q_{x-1}$ the colour is chosen, among all colours already occurring within this particular level $x-1$ structure, proportionally to the number of level x structures in which they appear; whereas with probability q_{x-1} the colour is chosen according to a superior level. In this second scenario, we move to the superior level and then repeat the operation (with q_{x-i} using $i = 2$ then higher values if necessary) until one of two situations is encountered. If the colour has to be chosen within a level y structure, which occurs with probability $(1-q_y)\prod_{z=y-1}^{x-1} q_z$, the colour is chosen among all colours occurring in this level y structure, proportionally to the number of level $y+1$ structures in which they appear. If the colour has to be chosen according to a level superior to level 1, a new colour is introduced with probability q_0 or chosen in the other level 1 structures with probability $1-q_0$. Thus, for an HPA process with d levels, the needed parameters are p_i with $i \in [1,d]$ and q_j with $j \in [0, d-1]$. Some trivial parameters that will be useful are $p_0 = 0$ and $q_d = p_{d+1} = 1$; respectively meaning we never create a new system, never put a ball twice in the same urn, nor put balls within balls.

Table 6.1 – **Summary of mathematical quantities involved in HPA**

$B_k^{(S)}$	Probability of a new structure appearing at level k during a time step
$G_k^{(S)}$	Probability of a level k structure getting a new level $k+1$ structure
$B_k^{(N)}$	Probability of a new ball colour appearing at level k, $B_k^{(N)} = Q \; \forall \; k$
$G_k^{(N)}$	Probability of a ball colour appearing in a new level k structure
Q	Probability of a new colour/node being born in a time step, $B_k^{(N)} = Q \; \forall \; k$
$S_{k,n}$	Number of level k structures containing n level $k+1$ structures
$N_{k,m}$	Number of ball colours appearing in m level k structures
P_k	Probability that a level $k+1$ structure belongs to a level k structure of size one
R_k	Probability that the colour of the ball involved in an event was chosen according to level k

We then map these construction rules onto an embedded systems of preferential attachment equations. For each Δt, the structures of level k have probabilities of birth, $B_k^{(S)}$, and growth, $G_k^{(S)}$, given by

$$B_k^{(S)} = \sum_{i=1}^{k} p_i \prod_{j=1}^{i-1} (1-p_j) \quad \text{and} \quad G_k^{(S)} = p_{k+1} \prod_{i=1}^{k} (1-p_i) \tag{6.3}$$

since birth events occur if structures are created at level k or at a superior level ($k' < k$), but growth events require a creation at level $k+1$. From these probabilities, the number $S_{k,n}(t)$ of structures of level k with size n can be approximately followed using Eq. (6.1) (a more complete set of embedded equations is given in Sec. 6.5):

$$S_{k,n}(t+\Delta t) = S_{k,n}(t) + B_k^{(S)}\delta_{n,1} + \frac{G_k^{(S)}}{\left(G_k^{(S)} + B_k^{(S)}\right)t}\left[(n-1)\,S_{k,n-1}(t) - nS_{k,n}(t)\right] \ . \tag{6.4}$$

From Eq. (6.2), we obtain the asymptotic scaling of structure sizes at level k:

$$\gamma_k^{(S)} = 2 + \frac{B_k^{(S)}}{G_k^{(S)}} \ . \tag{6.5}$$

While the description of structure sizes is a straightforward problem, things get convoluted for the number $N_{k,m}(t)$ of colours appearing in m structures of level k. An important logical constraint occurs for structures of level $k < d$ with size equal to one: if the colour is new for its sole inferior structure of level $k+1$, it must logically be new for the structure of level k. Thus, the probabilities $\{q_k\}$ are not strictly respected, but instead follows a corrected set of probabilities:

$$q'_k(t) = q_k + P_{k-1}(t)\left(1-q_k\right) = q_k + \frac{S_{k-1,1}(t)}{\sum_n nS_{k-1,n}(t)}\left(1-q_k\right) \tag{6.6}$$

where $P_{k-1}(t)$ is the probability that the structure of interest at level k is the sole substructure of the selected structure of level $k-1$. In other words, if the colour is new at level k, it can either be because of the initial probability q_k, or because it was forced to be new by the aforementioned logical constraint. The probabilities $P_{k-1}(t)$ can be obtained from the master equation for sizes of level $k-1$ structures, and so can their steady state values in the limit $t \to \infty$ (as per our usual method, see Sec. 6.5). These values yield

$$\lim_{t\to\infty} q'_k(t) = q_k + \frac{(1-q_k)}{1+2G_{k-1}^{(S)}/B_{k-1}^{(S)}} \ . \tag{6.7}$$

It is then a matter of evaluating the birth and growth probabilities, $B_k^{(N)}$ and $G_k^{(N)}$ for colours occurrences within level k. As a new colour must simultaneously appear in one structure of all levels, the birth probabilities are simply given by the global birth probability Q (total probability that a new colour is introduced during a time step):

$$B_k^{(N)} = Q = \sum_{x=1}^{d+1} p_x \left[\prod_{y=0}^{x-1} q'_y\right]\left[\prod_{z=0}^{x-1}\left(1-p_z\right)\right] \quad \forall\, k \ . \tag{6.8}$$

To obtain the growth probabilities, we first write the probability R_k that the chosen colour is an existing one selected according to level k. These probabilities are easily calculated for the two lowest structural levels (e.g., R_d implies that we reach level d and select an existing colour at level $d-1$):

$$R_d = \left(1-q'_{d-1}\right)\prod_{i=1}^{d-1}\left(1-p_i\right) \tag{6.9}$$

$$R_{d-1} = p_{d-1}\left(1-q'_{d-2}\right)\prod_{i=1}^{d-2}\left(1-p_i\right) + q'_{d-1}\left(1-q'_{d-2}\right)\prod_{i=1}^{d-1}\left(1-p_i\right) \ . \tag{6.10}$$

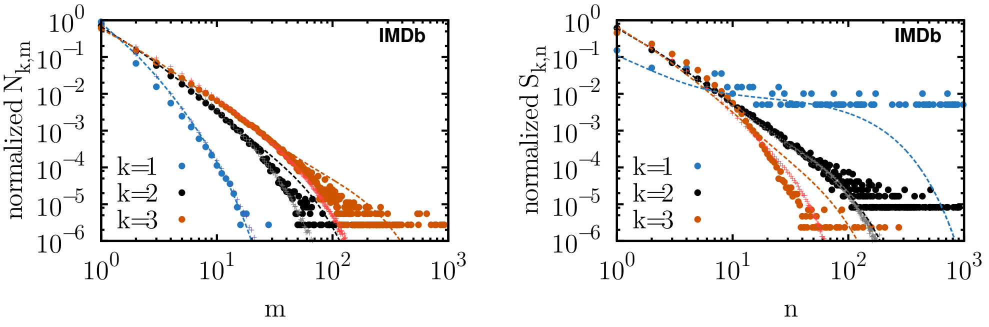


Figure 6.2 – **Hierarchical structure of movie production.** Events involving producers are distributed among three structural levels: movies on level 1, production companies on level 2 and countries on level 3. (left) Distribution of the number of movies/companies/countries a given producer is involved with. (right) Number of producers/movies/companies involved within a given movie/company/country. The empirical data is shown in dots, while lines and crosses are obtained by iterating Eqs. (6.4) and (6.13) and by direct Monte-Carlo simulation of HPA respectively for 10^6 time steps using $p_1 = 0.0005$, $p_2 = 0.185$, $p_3 = 0.385$, $q_1 = 0.80$, $q_2 = 0.60$, and $q_3 = 0.50$. Simulated results of $S_{1,n}$ are not shown to avoid cluttering the figure (note that the plateau observed in the empirical data is due to finite size). The correspondence between the observed scale exponents and our mathematical results implies that the model is not over parametrized. The chosen parameters were selected to roughly reproduce the qualitative behaviour of each distribution.

Extrapolating from the construction of these probabilities yields a recursive expression:

$$R_{k-1} = p_{k-1}\left(1-q'_{k-2}\right)\prod_{i=1}^{k-2}\left(1-p_i\right) + q'_{k-1}\left(1-q'_{k-2}\right)\frac{R_k}{1-q'_{k-1}}\ , \tag{6.11}$$

starting from R_d given above. The terms $G_k^{(N)}$ can then be written as the sum of the probabilities of choosing an existing node according to level k or any superior level:

$$G_k^{(N)} = \sum_{i=1}^{k} R_i\ . \tag{6.12}$$

This result respects $Q(t) + G_k^{(N)}(t) = 1$ for $k = d$ as we always add at least one ball to one urn. We here made the time dependency explicit since $\{q'_i\}$ are time dependent for small t. It is finally obtained that the numbers of colour appearing m times in level k follow

$$N_{k,m}(t+\Delta t) = N_{k,m}(t) + Q(t)\delta_{m,1} + \frac{G_k^{(N)}(t)}{\left[Q(t)+G_k^{(N)}(t)\right]t}\left[\left(m-1\right)N_{k,m-1}(t) - mN_{k,m}(t)\right]\ . \tag{6.13}$$

where we make the time dependency of $Q(t)$ and $G_k^{(N)}(t)$ explicit, since $\{q'_i\}$ are time dependent for small t. For $t \to \infty$, the system converges to statistical equilibrium and scales as $m^{-\gamma_k^{(N)}}$ following $\{q'_i\}$ given by Eq. (6.7) and:

$$\gamma_k^{(N)} = 2 + \frac{Q}{G_k^{(N)}}\ . \tag{6.14}$$

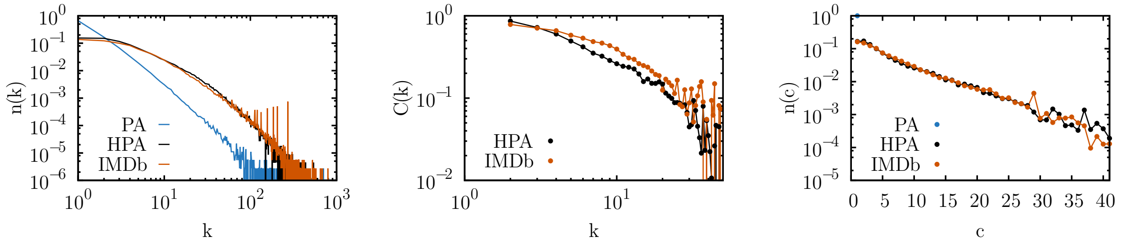


Figure 6.3 – **Scale independence and clustering of a projected hierarchical systems.** (left) Degree distribution observed in networks created by projecting the systems of Fig. 6.2 on webs of co-production credits (the actual data and one simulated system). A network obtained through the classic preferential attachment model [12] (PA) is given for comparison. (middle) Average clustering coefficient for nodes of degree k. PA leads to a vanishing clustering $C(k) = 0$ for all degree k in large networks. (right) Distribution of node centrality measured with their coreness c under k-core decomposition of the networks. PA leads to a unique shell of coreness $c = 1$ because of the network's tree like structure.

To validate Eqs. (6.4) and (6.13), we reproduce the pyramid of production entities in the movie industry. Based on the Internet Movie Database (IMDb), we study a system with 3 structural levels where producers (coloured balls: one ball is one producing credit, while a colour represents a unique producer) are thrown in films (urns) which are associated with one principal production company (bins), itself associated with one country (boxes). The results of this case study are presented in Fig. 6.2. To fix the HPA parameters with the empirical data, we simply start by reproducing by eye the size distributions from the top down since they are independent of $\{q_i\}$ and only depend on the size of the superior levels. We then reproduce the membership distributions from the bottom up since they only depend on the previously fixed $\{p_i\}$ and the memberships at inferior levels.

6.3 Complex networks as an emerging property of HPA

Even with the advent of large databases, few hierarchical systems are categorized and referenced as such. Consequently, research tend to focus on a single level of activity. For instance, the IMDb is often studied as a network of co-actors [12, 98]. Or, in the case of the system used in Fig. 6.2., a network of co-production where producers are connected if they produced a film together (if their balls are found within a common level d structure). Effectively, this implies that the system is reduced to a projection of all structural levels onto the chosen activity. While the involvement of actors and producers in movies are well captured, their involvement in different companies and countries is then encoded, and thus usually lost, in the structure of the resulting network.

Figure 6.3 presents some basic properties obtained by projecting the movie production hierarchical system onto a network of co-producing credits. Namely, we first investigate the degree distribution $n(k)$ (co-producing link per producer) and the clustering function $C(k)$ (probability that two links of a degree k producer are part of a triangle) of a network projection of a HPA system based on the parameters used in Fig. 6.2. The non-trivial clustering [98, 90] and the power-law tail of the degree distribution [12], properties ubiquitous in real networks, are reproduced here as emergent features of our HPA model. Similarly, Fig. 6.3 also presents result of a centrality analysis known as core decomposition. This analysis relies on the concept of c-cores, i.e., the maximal subset where all nodes share c links amongst one another. A node is consequently assigned the coreness c if it belongs to the c-core but not to the $(c+1)$-core. This procedures effectively defines a periphery (low c) and core (high c) to the network and was recently shown to reflect structural organization beyond mere local correlations [59]. Consequently, we can be confident that the model effectively reproduces the structure of the real hierarchical systems beyond the statistical properties previously considered in Fig. 6.2.

Besides scale-independent degree distribution and non-trivial clustering function, the fractality of complex networks is often a tell-tale sign of hierarchical organization [94, 95]. One can unravel the fractal nature of a network using a box counting method: groups of nodes within a distance (number of links) r of each other are grouped assigned to the same box. The fractal dimension d_b of a network manifests itself as a scaling relation between the number N_b of boxes needed to cover all nodes and the size r of the boxes ($N_b \propto r^{-d_b}$). The self-similarity of network structure was previously assumed to stem from a repulsion or disassortativity between the most connected nodes [95]. However, Fig. 6.4 demonstrates that fractality can also emerge from a scale-independent hierarchical structure, without further assumptions. Interestingly, Fig. 6.4(left) also illustrates how, even if fractality might imply hierarchy, the opposite is not necessarily true.

The box decomposition method tells us something about how networks cover the space in which they are embedded, and consequently at what speed a walker might encounter new nodes in this network. However, it tells us nothing about the geometrical space that supports the network, or how a walker could find one specific node. In that respect, the navigability of complex networks has recently been a subject of interest for two reasons. First, the development of a mapping of networks to a geometrical space allows to predict the probability of links as a function of geometrical distance between nodes, which in turn enables an efficient navigation through the network [20, 21]. Second, network growth based on preferential attachment fails to capture this geometrical property [84]. In a recent paper [84], this metric was consequently considered as evidence of an opposition between two organizational forces: popularity (preferential attachment) and similarity (assortativity). Our last case study, shown in Fig. 6.4(right), indicates that geometrical constraints, or network navigability, can emerge under a strict preferential attachment; which implies a growth driven by popularity only, but one occurring on multiple structural levels. The different hierarchical levels can *a posteriori* be interpreted as indicators of similarity, but are conceptually much more general.

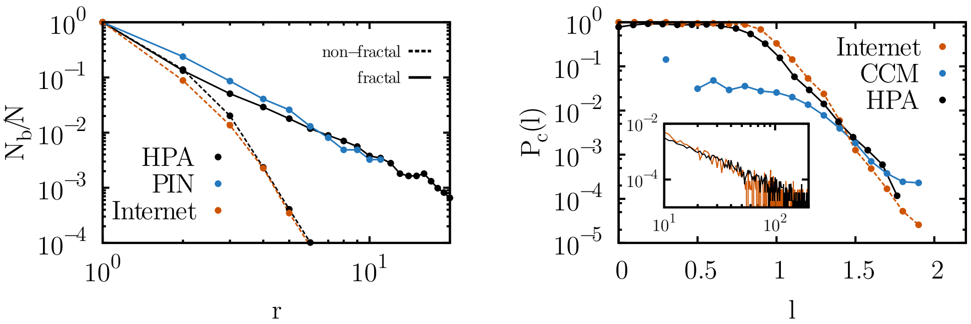


Figure 6.4 – **Fractality and navigability of projected hierarchical systems.** (left) Box counting results on a well-known fractal network (protein interaction network of Homo Sapiens) and a non-fractal network (the Internet at the level of autonomous systems). HPA can approximately model how both of these networks span and cover their respective space, with $p_1 = 0.01$, $p_2 = 0.02$, $p_3 = 0.30$, $q_1 = 0.95$, $q_2 = 0.80$ and $q_3 = 0.30$ (fractal) or $p_1 = 0.005$, $p_2 = 0.195$, $p_3 = 0.395$, $q_1 = 0.60$, $q_2 = 0.40$ and $q_3 = 0.30$ (non-fractal). The conditions in which HPA leads to fractality are discussed in Sec. 6.5. (right) Probability of connection $P_c(l)$ between nodes at a distance l after an inferred projection of the networks onto an hyperbolic space. (The distance is given as a fraction of the hyperbolic disc radius. See Boguñá *et al.* [21] or Sec. 6.6 for details on the method.) Both the Internet and its HPA model are the same as presented on the left and share a similar scaling exponent for their degree distribution (see inset: distribution p_k versus degree k). The CCM corresponds to a rewired network preserving degree distribution and degree-degree correlations, but obviously lacking the more complex structural correlations.

We also compare the results obtained on the actual network and on its HPA model with those obtained on a rewired network that preserves the degree distribution and degree-degree correlations (Correlated Configuration Model) [76]. The fact that this last process does not preserve the navigability of the Internet structure indicates that it emerges mostly from long-range correlations. As the HPA network does reproduce the navigability of the Internet, these long-range correlations could very well be consequences of the hierarchical structure. It would thus be instructive to investigate whether the inferred structure corresponds to the actual hierarchy of the Internet (probably of geographical nature: continents, countries, regions).

6.4 Conclusion

This chapter is a proof of concept for the Hierarchical Preferential Attachment model (HPA) which reproduces the hierarchical nature of complex systems.We have illustrated with case studies how complex networks are better reproduced by first modelling their hierarchical structure, and then projecting this structure onto a network. Not only does this procedure yields the non-trivial clustering of networks and their degree/centrality distributions at multiple levels, but it also gives access to the hidden geometrical metrics of these networks and the way they occupy space.

The fact that so many key features of network structure are modelled using two minimal assumptions, hierarchy and preferential attachment, indicates that HPA provides more than theoretical insights; it leads credence to these assumptions. HPA could therefore be used to infer the possible hierarchical structure of networks where this information is not directly available.

Finally, while HPA is essentially a simple stochastic growth process, it is one which perfectly illustrates how complex structural features of real networks — e.g. scale independence, clustering, self-similarity, fractality and navigability — can emerge through the hierarchical embedding of scale-independent levels. Perhaps this is the most important lesson here: to study the structure of complex networks, one should avoid focusing on unique level of activity (e.g. links), but instead investigate the hidden hierarchical organizations from which the networks emerge.

6.5 Appendix 6.A: Notes on HPA

6.5.1 Reminder: solving a preferential attachment process

We study any process where the evolution of the number of elements $N_s(t)$ with share s at time t follow a master equation of the type:

$$N_s(t+\Delta t) = N_s(t) + B\delta_{s,1} + \frac{G}{(B+G)\,t}\left[(s-1)\,N_{s-1}(t) - sN_s(t)\right] \ . \tag{6.15}$$

Since B is the birth probability, the evolution of the normalized distribution $\{n_s(t)\}$ can be obtained by replacing $N_s(t)$ by $Btn_s(t)$:

$$B\,(t+\Delta t)\,n_s(t+\Delta t) = Btn_s(t) + B\delta_{s,1} + \frac{GB}{B+G}\left[(s-1)\,n_{s-1}(t) - sn_s(t)\right] \ . \tag{6.16}$$

Since Δt is an arbitrary time step, and B and G are just as arbitrary, we can use an equivalent process in continuous time by using $\Delta t \to dt$,

$$B\,(t+dt)\,n_s(t+dt) = Btn_s(t) + dt\left\{\frac{GB}{B+G}\left[(s-1)\,n_{s-1}(t) - sn_s(t)\right] + B\delta_{s,1}\right\} \ , \tag{6.17}$$

from which the following set of ordinary differential equations is obtained:

$$\lim_{dt\to 0}\frac{(t+dt)\,n_s(t+dt) - tn_s(t)}{dt} = \frac{d}{dt}\left[tn_s(t)\right] = \frac{G}{B+G}\left[(s-1)\,n_{s-1}(t) - sn_s(t)\right] + \delta_{s,1} \ . \tag{6.18}$$

Solving at statistical equilibrium, i.e., $n_s(t) = n_s^*$ such that $\frac{d}{dt}n_s(t) = 0 \ \forall \ s$, yields

$$\left(1 + s\frac{G}{B+G}\right)n_s^* = \frac{G}{B+G}\,(s-1)\,n_{s-1}^* + \delta_{s,1} \tag{6.19}$$

or more directly

$$n_s^* = \frac{\prod_{m=1}^{s-1} \frac{G}{B+G} m}{\prod_{m=1}^{s} \left(1 + m\frac{G}{B+G}\right)} \; . \tag{6.20}$$

Assuming a scale-free behaviour for $s \to \infty$, this steady state can be shown to scale as

$$\lim_{s\to\infty} \frac{n_{s+1}^*}{n_s^*} = \left(\frac{s+1}{s}\right)^{-\gamma} \qquad \text{where } \gamma = 2 + \frac{B}{G} \; . \tag{6.21}$$

6.5.2 On structure sizes in HPA

In Sec. 6.2.3, structure sizes were approximated to follow simple equations under the general form of Eq. (6.15). However, while this is exact for the first structural level, the probability that a structure of size n in level 2 will depend on the size m of the level 1 structure in which it is nested. Mathematically, level 1 evolves according to

$$S_{1,m}(t+\Delta t) = S_{1,m}(t) + p_1\delta_{m,1} + \frac{(1-p_1)\,p_2}{[p_1 + (1-p_1)\,p_2]\,t}\left[(m-1)\,S_{1,m-1}(t) - mS_{1,m}(t)\right] \tag{6.22}$$

and level 2 follows

$$\begin{aligned} S_{2,n,m}(t+\Delta t) &= S_{2,n,m}(t) + p_1\delta_{n,1}\delta_{m,1} \\ &+ \frac{(1-p_1)\,mS_{1,m}(t)}{[p_1+(1-p_1)p_2]\,t}\left\{(1-p_2)p_3\frac{(n-1)S_{2,n-1,m}(t) - nS_{2,n,m}(t)}{\sum_i iS_{2,i,m}(t)} - p_2\frac{S_{2,n,m}(t)}{S_{1,m}(t)}\right\} \\ &+ \frac{(1-p_1)p_2(m-1)S_{1,m-1}(t)}{[p_1+(1-p_1)p_2]\,t}\left\{\frac{S_{2,m-1,n-1}(t)}{S_{1,m-1}(t)} + \delta_{n,1}\right\} \end{aligned} \tag{6.23}$$

where $S_{2,n,m}(t)$ is the number of level 2 structure which are of size n and nested in level 1 structure of size m.

To obtain a master equation following the form of Eq. (6.15) one must sum over all m while assuming $\langle n\rangle_{2,m} = \sum_i iS_{2,i,m}(t)/mS_{1,m}(t) = \langle n\rangle_2 \; \forall \; m$, such that $[p_1 + (1-p_1)p_2]\,t \cdot \langle n\rangle_2$ will be equal to $[p_1 + (1-p_1)p_2 + (1-p_1)(1-p_2)p_3]\,t$; yielding the form presented directly in Sec. 6.2.3 by considering uncorrelated levels of organization. These approximations are the source of the error observed in the mean-field description of distribution $S_{k,n}(t)$ for $k > 1$ and are progressively worse for higher k (lower structural levels). The progression in error is essentially caused by the fact that a strict description of the third level, for instance, should not only be given in terms of $S_{3,n,m}(t)$, but of $S_{3,n,m,l}(t)$ describing the number of level 3 structures of size n nested in level 2 structures of size m themselves nested in level 1 structures of size l.

6.5.3 On node memberships in HPA

We are interested in obtaining the thermodynamic limit, i.e., when $t \to \infty$, for the distribution of node memberships at level k (the distribution of level k structures in which a given color appears). In the main text, we solved the problem of logical constraints on node memberships by introducing a biased set of probabilities $\{q_i'\}$:

$$q_k'(t) = q_k + \frac{S_{k-1,1}(t)}{\sum_n n \cdot S_{k-1,n}(t)} \left(1 - q_k\right) \ . \tag{6.24}$$

Within our mean-field description of time evolution, these probabilities can be explicitly calculated at each time step using the current size distribution at level $k-1$. In the thermodynamic limit, we can instead use the steady-state distributions described by Eq. (6.20). Doing so yields

$$q_k' = q_k + \frac{S^*_{k-1,1}}{\sum_n n \cdot S^*_{k-1,n}} \left(1 - q_k\right) = q_k + \frac{G^{(S)}_{k-1} + B^{(S)}_{k-1}}{2G^{(S)}_{k-1} + B^{(S)}_{k-1}} \frac{1 - q_k}{\langle n \rangle_{k-1}} \tag{6.25}$$

where the average size $\langle n \rangle_{k-1}$ is straightforwardly calculated as the ratio of total events to birth events:

$$\langle n \rangle_{k-1} = \frac{\left[G^{(S)}_{k-1} + B^{(S)}_{k-1}\right] t}{\left(B^{(S)}_{k-1}\right) t} \tag{6.26}$$

such that

$$q_k' = q_k + \frac{\left(1 - q_k\right) B^{(S)}_{k-1}}{2G^{(S)}_{k-1} + B^{(S)}_{k-1}} \ . \tag{6.27}$$

This set of probabilities is then used to obtain master equations following Eq. (6.15) by assuming that memberships and structure sizes are uncorrelated. We here assume that nodes with m and m' memberships at level k see the same structure size distribution at level $k+1$. Such that an event at level $k+1$ is m/m' times more likely to involve the first node and not the latter. Similarly, when a structure is created at level $k-1$, we are effectively adding a membership at level k. Assuming that the node with m memberships at level k has m/m' more memberships at level $k-1$ than the node with m' memberships, these events affect the distribution in the same manner than regular growth events.

6.5.4 Multiple scale independence

We showed how we can only approximately follow the time evolution of the size distributions. However, we can derive multiple scale exponents in the thermodynamic limit $t \to \infty$. In the main text, we found the general scaling exponent for the size distribution of level k structures and, using the steady-state value of the corrected probabilities, we also found the scaling exponent for node memberships.

When looking at projected properties of a hierarchical system, for instance the degree distribution of the resulting network, we can compose the membership and size distributions of the lowest level (where links are created) and deduce the resulting scaling exponent. As shown in [57], the idea is to define the following *probability generating functions* (pgf):

$$\mathcal{S}(x,t) = \sum_n S_{d,n}(t)x^n \qquad \text{and} \qquad \mathcal{N}(x,t) = \sum_m N_{d,m}(t)x^m \; . \tag{6.28}$$

As a community of size n implies $n-1$ links for each node, the first of these distributions can be generated by

$$\mathcal{S}_l(x,t) \equiv \frac{\frac{d}{dx}\mathcal{S}(x,t)}{\frac{d}{dx}\mathcal{S}(x,t)|_{x=1}} = \frac{\sum_n S_{d,n}(t)nx^{n-1}}{\sum_n S_{d,n}(t)n} \; . \tag{6.29}$$

The degree distribution is then generated by a pgf $\mathcal{D}(x,t)$ combining the distribution of memberships and that of links obtained from each of this membership:

$$\mathcal{D}(x,t) = \mathcal{N}(\mathcal{S}_l(x,t),t) \; , \tag{6.30}$$

which simply scales as the slowest falling function between $\mathcal{N}(x,t)$ (of scale exponent $\gamma_d^{(N)}$) and $\mathcal{S}_l(x,t)$ (of scale exponent $\gamma_d^{(S)} - 1$ because of the derivative in Eq. 6.29). The scale exponent of the degree distribution is thus given by

$$\min\left[\gamma_d^{(N)}, \gamma_d^{(S)} - 1\right] \; . \tag{6.31}$$

The same method could of course be used to determine the scaling of other projections; e.g., network of companies sharing or having shared at least one producer.

6.5.5 On the fractality of projected networks

As previously shown, HPA can produce both fractal and non-fractal networks. Since the definition of network fractality is somewhat ambiguous, so is the distinction between set of HPA parameters leading to fractality or not. However, a rule of thumb can be established.

The dimensionality of a network is analysed through standard box-counting: boxes of sizes r cover groups of nodes all within a distance $r-1$ or less of one another (the distance being measured in number of links). Fractal networks are characterized by a box-counting dimension d_b governing the number $N_b(r)$ of boxes of size r needed to cover all nodes. This relation takes the form $N_b(r) \propto r^{-d_b}$. It remains to be determined whether or not this box counting method is truly equivalent to an actual measure of *dimensionality*. Nevertheless, it can at the very least be interpreted as an observation of how easily a network can be covered. This in itself is interesting and is similar to a spreading process, or other dynamics which can be coarse grained.

Most models of stochastic network growth produce networks with very low mean shortest paths, low clustering and no long-range correlations. Consequently, the number of boxes needed to cover the whole network decreases very rapidly. In HPA, we can control the way

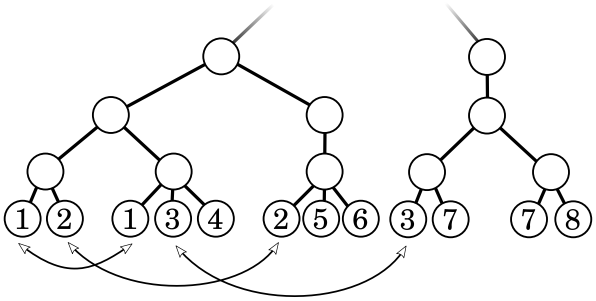


Figure 6.5 – **Example of how nodes act as bridges between structures in HPA.**

Table 6.2 – **Parameters used in the case study of fractal behaviour in HPA.**

Network	p	q
A	[0.1, 0.2, 0.3]	[0.69, 0.9, 0.1]
B	[0.01, 0.02, 0.3]	[0.95, 0.8, 0.3]

boxes cover the network since the distance between higher structural level is directly influenced by the memberships at this level. Hence, HPA can generate networks that are robust to box covering (i.e., such that $N_b(r)$ decreases slower with regards to r) if higher structural levels feature less nodes that act as bridges between structures and levels. For example, in Fig. 6.5, only nodes 1, 2 and 3 can be used by boxes to move from one level to the other (from cities to countries, here illustrated as an inverted tree).

More precisely, let us consider two different networks: A and B, built using the parameters given in Table 6.2. Roughly speaking, in network A, level 2 structures contain on average two level 3 structures whereas nodes belong to over 4 level 3 structures. Therefore, a single node typically grants access to all level 3 structures contained within any of its level 2 structure, such that a box covering at least *part* of a level 2 structure typically covers *all* of it. The network is thus easily invaded as higher levels are not any harder to navigate.

In contrast, level 2 structures of network B contain on average eleven level 3 structures. An average node may be found within four level 3 structures, so that even a single average level 2 structure may have nodes at a distance greater than three steps. The network is thus harder to cover and can be expected to be much more robust to box-covering. As a general rule of thumb, we found that to feature *measurable* network self-similarity the average size of a structure (at level x) had to be at least greater than the memberships of a node at the lower level (at level $x + 1$).

6.6 Appendix 6.B: Data sets and network analysis algorithms

6.6.1 Data

Co-producers in the Internet Movie Database All information related to the production of movies were downloaded from the Internet Movie Database in August 2013. The data sets, available as text files are as follows.

- For a given movie, the main country of production is listed;
- for a given movie, the main production company is listed along with secondary companies identified with, e.g., “in association with” or “with the participation of”;
- for each producer in the database, all productions in which he or she was involved is listed.

Any movie which could not be cross-referenced across the three lists was removed from the dataset. For the rest, we obtained a list of producers involved in movies produced by companies (main production company only) in given countries. This list contains 363,571 producers involved in 42,691 movies produced by 12,196 companies in 197 countries,

Protein interaction network of Homo Sapiens The protein interaction network of Homo Sapiens was downloaded from the DIP database available at
`http://dip.doe-mbi.ucla.edu/dip/Main.cgi`.

According to their homepage, “the DIP^{TM} database catalogues experimentally determined interactions between proteins. It combines information from a variety of sources to create a single, consistent set of protein-protein interactions. The data stored within the DIP database were curated, both, manually by expert curators and also automatically using computational approaches that utilize the the knowledge about the protein-protein interaction networks extracted from the most reliable, core subset of the DIP data.”

Internet This dataset is a symmetrized snapshot of the structure of the Internet at the level of autonomous systems, reconstructed from BGP tables posted at *archive.routeviews.org*. This snapshot was created by Mark Newman from data for July 22nd 2006. The network contains 22 962 nodes and 48 436 links.

6.6.2 Fractal dimension through box counting

To investigate how networks cover space, we use the box counting algorithm of Song *et al.* [94]. The algorithm is simple. For a given box size r,

- i. select an unassigned node k at random as a seed;
- ii. assign one of the maximal (non-unique) subset of nodes that are within a distance $r - 1$ of k and of one another to a box;
- iii. repeat until all nodes are assigned to a box.

We are then interested in how the number of boxes needed to cover the network scales with regards to box size r. As explained by Song *et al.* [94], while their are multiple covering possibles, they lead to the same scaling exponents (or lack thereof).

6.6.3 Centrality structure via k-core decomposition

In this chapter, we use a measure of *centrality* based on the concept of coreness. A node's coreness c is defined through its place in the network's k-core decomposition. This decomposition separates the network into different shells, thereby identifying the effective *core* and *periphery* of the network. A k-core is a subset of the original network where all nodes included have degree equal or higher than k amongst one another. We then define the k-shells (the ensemble of nodes of coreness $c = k$) as the nodes part of the k-core, but not of the $(k + 1)$-core.

Through this definition, the coreness may appear complex to compute, but a simple algorithm allows us to do the decomposition very efficiently [15].

```
**Input** graph as lists of nodes $\mathcal{V}$ and neighbours $\mathcal{N}$
**Output** list $\mathcal{C}$ with coreness for each node
compute and list the degrees $\mathcal{D}$ of nodes;
order the set of nodes $\mathcal{V}$ in increasing order degrees;
**for all** $v \in \mathcal{V}$ in the order **do**
  $\mathcal{C}(v) := \mathcal{D}(v)$;
  **for all** $u \in \mathcal{N}(v)$ **do**
    **if** $\mathcal{D}(u) > \mathcal{D}(v)$ **then**
      $\mathcal{D}(u) := \mathcal{D}(u) - 1$;
      Reorder $\mathcal{V}$ accordingly
    **end if**
  **end for**
**end for**
```

In a recent publication, we show how the coreness structure can be used as an efficient analytical model for percolation on networks; reproducing degree-degree correlations and unique features of power-grids for instance [59].

6.6.4 Mapping to a hyperbolic space

The problem of network navigability lies in the apparent paradox between their complex structure and the efficiency with which they seem to be navigable. Be it how people are able to identify potential paths between them and a stranger based only on local information and some identifying traits of the stranger [97]; or how fast routing is achieved on the Internet without a global map [21]. This begs the question, is it possible to obtain the underlying map in which we seem to navigate? If so, in what kind of space are these networks embedded?

We use a recent algorithm that maps complex networks to hyperbolic space [84]. As each point of this hyperbolic space is a saddle point, the algorithm can use the local manifolds to account for how an individual can be part of different social groups that have nothing to do with each other: groups on different manifolds may be close to the same node, but do not have to be close to one another.

While we avoid going into too much details here, let us just paint a broad picture of how the algorithm attempts to draw the map of a given network. Now that we chose a space in which to draw the network (hyperbolic), we must assign a position (r_i, θ_i) to each node $i \in [1, N]$ in a way that maximizes the likelihood [1] of the observed links as a function of all node positions. With observations (i.e., number of links) typically scaling as N, and inferred parameters scaling as $2N$, this is not a simple matter. To simplify the problem, the algorithm fixes the radial position of a node as a decreasing function of its degree; the idea being that the nodes with more links are more likely to end up closer to everyone else and thus near center. The nodes are then separated into layers, the number of which is an input parameter that needs to be varied to assure robust results. The angular position of each node within a layer is then optimized iteratively: starting by optimizing θ_i given the other θs in the layer (marginal distribution), and iterating through all nodes more than once (until convergence, usually of the order of maximal degree within the layer).

When we have a map, i.e., a position (r_i, θ_i) in space for each node, we can measure the distribution of hyperbolic distance l between nodes and the probability $P_c(l)$ that nodes at a distance l are connected. On random networks, the algorithm tend to separate *all* nodes by some minimal distance; such that nodes are never *near* one another. This is because the following statement rarely holds within random network: node i is close to node j and k, so node j and k must be close. Even with clustering, their other neighbours would tend to be far apart. In fact, this is what we observe using HPA with $d = 1$ (SPA). The way to enforce some geometrical-like constraints in random networks is to consider a hierarchical structure. In our case, a hierarchical dimensionality $d = 3$ appears to be sufficient.

[1] As in any Bayesian inference problem, we need to fix a model that describes the likelihood of an observation (e.g., link between i and j) as a function of our inferred parameters $(r_i, \theta_i, r_j, \theta_j)$. Without getting into the details, we obviously want the connection probability to decrease with hyperbolic distance.

Conclusion

My five years in graduate school were mostly inspired by the following quote from Herbert Simon [92]:

> Human beings, viewed as behaving systems, are quite simple. The apparent complexity of our behaviour over time is largely a reflection of the complexity of the environment in which we find ourselves.

Consequently, this thesis tackled the following questions: how well can we describe social systems based on simple principles and processes? In what context can human interactions be reduced to games of balls and urns? The first step to answer our questions was to define scale independence and universality; to suggest that the microscopic details of some social systems did not matter as the same lack of characteristic scale was found across very diverse activities and populations. These concepts led us to focus on a certain class of social systems which can be studied by statistical physics. Our goal was thus to develop the statistical physics of social systems and to describe how resources and activities are shared, or how those distributions and structures change in time, based on the knowledge of their asymptotic scale-independent state.

Statistical physics of social systems? Obviously, while we are still far from a true statistical physics of social systems, we have successfully applied statistical physics to study the evolution of resource distribution in social systems. In light of the work done, we can postulate this explanation to Herbert Simon's statement: human interactions are quite simple from a macroscopic point-of-view. We can describe statistical or structural properties of social systems with simple processes, as long as we do not zoom in on the microscopic details. In fact, most of the microscopic details are lost in our assumption of indistinguishability between individuals with the same past involvement in the system.

Similarly, in part because of this assumption, our description *does not* imply that we actually gain a mechanistic explanation for the growth of these systems. As was highlighted in Chap. 4, plenty of different null models with different assumptions and mechanisms happen to fall within our framework. To quote Herbert Simon again [92],

> No one supposes that there is any connexion between horse-kicks suffered by soldiers in the German army and blood cells on a microscope slide other than that the same urn scheme provides a satisfactory abstract model of both phenomena.

That being said, what did we learn about that urn scheme? How much can we learn by throwing balls in urns through the preferential attachment process? Is it just a toy model for the growth of scale-independent systems or can it actually be used to predict or infer properties and behaviour of actual complex systems?

Firstly, we saw that our simple description manages to capture some of the complex intricacies of social systems. Arbitrary growth rules do not reproduce the universal behaviours of some class of social systems, most importantly their scale independence. This suggests that some specific rules must be present to govern the emergence of said behaviours. For instance, there exists a self-organized coupling between the growth of a population as a function of the available resource, and the distribution of this resource among the existing population. This self-organization implies a set of constraints which, assuming that the underlying mechanism (i.e., $G(k)$) does not change in time, dictate the path that systems can take towards scale independence. These constraints effectively draw a map of how different systems reach their final organization. In essence, the knowledge of the state of a system at one point in time is sufficient to deduce where it came from and where it is going. Temporal data, seldom available in practice, can thus be extracted from a static snapshot of a resource distribution.

Secondly, we also discussed how the coupling of multiple scale-independent systems in a hierarchical fashion can lead to fascinating emergent behaviours. The basic idea is that social structures are a resource for individuals and, reciprocally, individuals are a resource for social structures. For example, scientists aim to collaborate with different institutions/labs or research groups just as these structures aim to grow and recruit scientists. This simple idea led us to generalize the preferential attachment game of balls and urns to an equivalent game of coloured balls in urns. The colours allow us to break, at least to some extent, the indistinguishability of different individuals and structures. We can now look at the structure of a system by tracking which colours (balls) are found in common urns (communities). Even with a single structural level of structure (urns), we can obtain a decent description of social networks with community structure. We thus gather some intuition on how modularity and clustering emerge in networks. The generalization of the process to a game with embedded levels of structure (urns in bins in boxes and so on) allows us to describe networks with more complex properties: hierarchical clustering, onion-like centrality structure, navigability, and self-similarity or fractality. In short, without directly considering network structure, but rather the hierarchical systems in which they live, we gain an understanding of how the complexity of complex networks can emerge.

Implications and applications Reducing the complex to the simple allows us to describe varied problems in the same language. The work presented in this thesis provides us with a simple and elegant dictionary to quantify and compare different systems by means of multiple structural and temporal scale exponents (γ and α). For more precision, we could consider the other temporal parameters: the non-linear deviation (κ), temporal delay (τ) and asymptotic state (b) for each relevant structural level (hierarchical dimension d).

Comparing systems of a different nature and size is a central question in network science. Which metrics allow us to efficiently differentiate, for example, a social network from a biological food web? Or a gene regulation network from a technological communication system? Talking about the dimensionality (d) of their relevant hierarchical structure is an interesting first step, be it an actual hierarchy or just an effective description. One could then zoom in to look at the different scale exponents of individual levels, or at their temporal description. For a given level, we find universality classes of growth and structure; just as for the whole system we could unify these in classes of classes. For instance, one could want to differentiate expanding hierarchies $\gamma_i > \gamma_{i+1}$ from contracting hierarchies $\gamma_i < \gamma_{i+1}$, or mixtures of both. These different mixing of levels and their effects on the lack of characteristic scale could be expected to lead to very different behaviours for the dynamical processes occurring within these structured systems.

This last step would allow the generalization of the framework developed here to more general problems. We could ask for instance what disease can invade a sexual network? When will a pathological disease be allowed to become endemic as the system evolves? Or how likely are public discussion forums to reach consensus on a given issue? What are their characteristic time scales to consensus if any? The hope is that by clarifying the coupling of temporal and structural features, we are also laying the groundwork toward a better understanding of social dynamics based on their underlying structure.

Finally, the path to scale independence remains to be further investigated. In a given system, what is intrinsic to its nature or to the considered activity? We saw in Chap. 3 that the growth functions for individuals involved in the Internet Movie Database and the Digg website remained relatively constant in time. Is this a feature shared by most social systems? In contrast, what is changing in time and self-organizing for criticality? And through what mechanisms does self-organization occur? Is it merely a question of entropic forces due to the distribution of resources and activities, or is a feedback loop taking place between social dynamics and social structure? In answering these questions, we can expect not only to broaden the applicability of the framework presented in this thesis, but also to better understand social systems and the relevance of criticality.

Bibliography

[1] M. Abramowitz and I. Stegun. *Handbook of Mathematical Functions.* Dover Publications, 1964.

[2] Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann. Link communities reveal multiscale complexity in networks. *Nature*, 466:761, 2010.

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

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

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

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

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

[8] S.K. Baek, S. Bernhardsson, and P. Minnhagen. Zipf's law unzipped. *New J. Phys.*, 13:043004, 2011.

[9] J. P. Bagrow, J. Sun, and D. ben Avraham. Phase transition in the rich-get-richer mechanism due to finite-size effects. *J. Phys. A: Math. Theor.*, 41:185001, 2008.

[10] P. Bak. *How Nature Works: the Science of Self-organized Criticality.* Springer, 1999.

[11] P. Bak, C. Tang, and K. Wiesenfeld. Self-Organized Criticality: An Explanation of 1/f Noise. *Phys. Rev. Lett.*, 59:381, 1987.

[12] A. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286:509, 1999.

[13] A. Barabási and R. Albert. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74:47, 2002.

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

[15] V. Batagelj and M. Zaveršnik. An O(m) algorithm for cores decomposition of networks. *arXiv:cs/0310049*, 2003.

[16] S. Bernhardsson, L. E. C. Da Rocha, and P. Minnhagen. The meta book and size-dependent properties of written language. *New J. Phys.*, 11:123015, 2009.

[17] S. Bernhardsson, L. E. C. Da Rocha, and P. Minnhagen. Size-dependent word frequencies and translational invariance of books. *Physica A*, 389:330, 2010.

[18] L. M. A. Bettencourt, J. Lobo, D. Helbing, C. Kuhnert, and G. B. West. Growth, innovation, scaling, and the pace of life in cities. *Proc. Natl. Acad. Sci. U.S.A.*, 104:7301, 2007.

[19] G. Bianconi and A.-L. Barabási. Competition and multiscaling in evolving networks. *Europhys. Lett*, 54:436, 2001.

[20] M. Boguná, D. Krioukov, and K. C. Claffy. Navigability of complex networks. *Nature Physics*, 5:74, 2009.

[21] M. Boguná, F. Papadopoulos, and D. Krioukov. Sustaining the internet with hyperbolic mapping. *Nature Communications*, 1, 2010.

[22] P. Want C. Song, T. Koren and A.L. Barabási. Modelling the scaling properties of human mobility. *Nature Physics*, 6:818, 2010.

[23] J. M. Carlson and J. Doyle. Highly optimized tolerance: Robustness and design in complex systems. *Phys. Rev. Lett.*, 84:2529, 2000.

[24] D. G. Champernowne. A model of income distribution. *Economic Journal*, 63:318, 1953.

[25] Y. Chang. What constitutes a normal speech? a study of the proportion of content words and function words in narration of chinese normal speakers. In *CCRE Aphasia poster presentation*.

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

[27] A. Clauset, C. Moore, and M. E. J. Newman. Hierarchical structure and the prediction of missing links in networks. *Nature*, 453:98, 2008.

[28] A. Clauset, C.R. Shalizi, and M.E.J. Newman. Power-law distributions in empirical data. *SIAM Rev.*, 51:661, 2009.

[29] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* John Wiley & Sons, 2012.

[30] L. E. C. Da Rocha, F. Liljeros, and P. Holme. Simulated Epidemics in an Empirical Spatiotemporal Network of 50,185 Sexual Contacts. *PLoS Comput Biol*, 7:e1001109, 2011.

[31] J. de Ruiter, G. Weston, and S. M. Lyon. Dunbar's Number: Group Size and Brain Physiology in Humans Reexamined. *American Anthropologist*, 113(4):557, 2011.

[32] D. de Solla Price. A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, 27:292, 1976.

[33] S. N. Dorogovtsev and J. F. F. Mendes. Evolution of networks with aging of sites. *Phys. Rev. E*, 62:1842, 2000.

[34] S. N. Dorogovtsev and J. F. F. Mendes. Evolution of networks. *Adv. Phys.*, 51:1079, 2002.

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

[36] S. N. Dorogovtsev, J. F. F. Mendes, and A. N. Samukhin. Structure of growing networks with preferential linking. *Phys. Rev. Lett.*, 85:4633, 2000.

[37] J. Doyle and J.M. Carlson. Power laws, highly optimized tolerance, and generalized source coding. *Phys. Rev. Lett.*, 84:5656, 2000.

[38] B. Drossel and F. Schwabl. Self-organised critical forest-fire model. *Phys. Rev. Lett.*, 69:1629, 1992.

[39] L. J. Dubé. Physique statistique avancée: notes de cours (2010).

[40] Robin Dunbar. Neocortex Size as a Constraint on Group Size in Primates. *Journal of Human Evolution*, 22(6):469, 1992.

[41] P. Erdős and A. Rényi. On random graphs. *Publicationes Mathematicae Debrecen*, 6:290, 1959.

[42] K. A. Eriksen and M. Hörnquist. Scale-free growing networks imply linear preferential attachment. *Phys. Rev. E*, 65:017102, 2001.

[43] J. W. Essam. Percolation theory. *Rep. Prog. Phys.*, 43:833, 1980.

[44] J. B. Estroup. *Gammes Sténographiques, 4e édition.* Institut Stenographique de France, Paris, 1916.

[45] T. Evans and R. Lambiotte. Line graphs, link partitions, and overlapping communities. *Physical Review E*, 80(1):016105, July 2009.

[46] M. Gerlach and E.G. Altmann. Stochastic model for the vocabulary growth in natural languages. *Phys. Rev. X*, 3:021006, 2013.

[47] R. Gibrat. *Les Inégalités économiques.* Librairie du Recueil Sirey, 1931.

[48] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proc. Natl. Acad. Sci.*, 99:7821, 2002.

[49] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proc. Natl. Acad. Sci. U.S.A.*, 99:7821, 2002.

[50] C. Godrèche and J. M. Luck. On leaders and condensates in a growing network. *J. Stat. Mech.*, P07031, 2010.

[51] B. Gonçalves, N. Perra, and A. Vespignani. Modeling Users' Activity on Twitter Networks: Validation of Dunbar's Number. *PLoS ONE*, 6(8):e22656, January 2011.

[52] R. Guimerà, L. Danon, A. Díaz-Guilera, F. Giralt, and A. Arenas. Self-similar community structure in a network of human interactions. *Phys. Rev. E*, 68:065103, 2003.

[53] B. Gutenberg and C. F. Richter. *Seismicity of the Earth and Associated Phenomena.* Princeton University Press, Princeton, New Jersey, 1949.

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

[55] L. Hébert-Dufresne, A. Allard, and L. J. Dubé. On the constrained growth of complex critical systems. *arXiv:1211.1361*, 2012.

[56] L. Hébert-Dufresne, A. Allard, V. Marceau, P.-A. Noël, and L. J. Dubé. Structural Preferential Attachement: Network organization beyond the link. *Phys. Rev. Lett.*, 107:158702, 2011.

[57] L. Hébert-Dufresne, A. Allard, V. Marceau, P.-A. Noël, and L. J. Dubé. Structural preferential attachment: Stochastic process for the growth of scale-free, modular, and self-similar systems. *Phys. Rev. E*, 85:026108, 2012.

[58] L. Hébert-Dufresne, A. Allard, J.-G. Young, and L. J. Dubé. Global efficiency of local immunization on complex networks. *Scientific Reports*, 3, 2013.

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

[60] L. Hébert-Dufresne, A. Allard, J.-G. Young, and L. J. Dubé. Universal growth constraints of human systems. *arXiv:1310.0112*, 2013.

[61] L. Hébert-Dufresne, E. Laurence, A. Allard, J.-G. Young, and L. J. Dubé. Complex networks are an emerging property of hierarchical preferential attachment. *arXiv:1312.0171*, 2013.

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

[63] L. Hébert-Dufresne, O. Patterson-Lomba, G. M. Goerg, and B. M. Althouse. Pathogen mutation modeled by competition between site and bond percolation. *Phys. Rev. Lett.*, 110:108103, 2013.

[64] B. Karrer and M. E. J. Newman. Random graphs containing arbitrary distributions of subgraphs. *Phys. Rev. E*, 82:066118, 2010.

[65] P. L. Krapivsky and S. Redner. Organization of growing random networks. *Phys. Rev. E*, 63:066123, 2001.

[66] P. L. Krapivsky and S. Redner. Statistics of changes in lead node in connectivity-driven networks. *Phys. Rev. Lett.*, 89:258703, 2002.

[67] P. L. Krapivsky, S. Redner, and F. Leyvraz. Connectivity of growing random networks. *Phys. Rev. Lett.*, 85:4629, 2000.

[68] Andrea Lancichinetti, Filippo Radicchi, José J Ramasco, and Santo Fortunato. Finding statistically significant communities in networks. *PLoS ONE*, 6(4):e18961, January 2011.

[69] Conrad Lee, Fergal Reid, Aaron Mcdaid, and Neil Hurley. Detecting Highly Overlapping Community Structure by Greedy Clique Expansion. 10, 2010.

[70] K. Lerman and R. Ghosh. Information Contagion: an Empirical Study of Spread of News on Digg and Twitter Social Networks. In *Proceedings of 4th International Conference on Weblogs and Social Media (ICWSM)*, 2010.

[71] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over Time: Densification Laws, Shrinking, Diameters and Possible Explanations. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2005.

[72] B. B. Mandelbrot. Towards a second stage of indeterminism in science. *Interdisciplinary Science Reviews*, 12:117, 1987.

[73] L. A. Meyers. Contact network epidemiology: bond percolation applied to infectious disease prediction and control. *Bull. Amer. Math. Soc.*, 44:63, 2007.

[74] B. R. Morin. Explicit solutions to the continuous time Albert-Barabási scale-free model. *arXiv:1105.0882*, 2011.

[75] S. Neale. *Hollywood Blockbusters: Historical Dimensions.* Routeledge, London, 2003.

[76] M. E. J. Newman. Assortative Mixing in Networks. *Phys. Rev. Lett.*, 89:208701, 2002.

[77] M. E. J. Newman. Spread of epidemic disease on networks. *Phys Rev E*, 66(1 Pt 2):016128, Jul 2002.

[78] M. E. J. Newman. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, 46:323, 2005.

[79] P. Minnhagen P, S. Bernhardsson, and B. J. Kim. Scale-freeness for networks as a degenerate ground state: a Hamiltonian formulation. *Europhys. Lett.*, 78:28004, 2007.

[80] G. Palla, A.-L. Barabási, and T. Vicsek. Quantifying social group evolution. *Nature*, 446:664, 2007.

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

[82] G. Palla, L. Lovász, and T. Vicsek. Multifractal network generator. *Proc. Natl. Acad. Sci. U.S.A.*, 107:7640, 2010.

[83] Gergely Palla, I.J. Farkas, P. Pollner, I. Derényi, and Tamás Vicsek. Fundamental statistical features and self-similar properties of tagged networks. *New Journal of Physics*, 10:123026, 2008.

[84] F. Papadopoulos, M. Kitsak, M. A. Serrano, M. Boguná, and D. Krioukov. Popularity versus similarity in growing networks. *Nature*, 489:537, 2012.

[85] V. Pareto. *Cours d'Économie Politique: Nouvelle édition par G.-H. Bousquet et G. Busino.* Librairie Droz, Geneva, 1964.

[86] J. W. Pennebaker. *The Secret Life of Pronouns: What Our Words Say About Us.* Bloomsbury Press, 2013.

[87] A. M. Petersen, J. N. Tenenbaum, S. Havlin, H. E. Stanley, and M. Perc. Languages cool as they expand: Allometric scaling and the decreasing need for new words. *Scientific Reports*, 2, 2012.

[88] W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery. *Numerical Recipes 3rd Edition: The Art of Scientific Computing.* Cambridge University Press, 2007.

[89] E. Ravasz and A.-L. Barabási. Hierarchical organization in complex networks. *Phys. Rev. E*, 67:026112, 2003.

[90] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabási. Hierarchical organization of modularity in metabolic networks. *Science*, 297:1551, 2002.

[91] D. Rybski, S. V. Buldyrev, S. Havlin, F. Liljeros, and H. A. Makse. Scaling laws of human interaction activity. *Proc. Natl. Acad. Sci.*, 106:12640, 2009.

[92] H. A. Simon. *Models of Man.* John Wiley & Sons, 1961.

[93] H. A. Simon. The architecture of complexity. *Proceedings of the American Philosophical Society*, 106:467, 1962.

[94] C. Song, S. Havlin, and H. A. Makse. Self-similarity of complex networks. *Nature*, 433:392, 2005.

[95] C. Song, S. Havlin, and H. A. Makse. Origins of fractality in the growth of complex networks. *Nature Physics*, 2:275, 2006.

[96] D. Sornette. *Critical Phenomena in Natural Sciences.* Springer, 2000.

[97] J. Travers and S. Milgram. An experimental study of the small world problem. *Sociometry*, 32, 1969.

[98] D. J. Watts and S. H. Strogatz. Collective dynamics of ’small-world’ networks. *Nature*, 393:440, 1998.

[99] J. C. Willis. *Age and Area: A Study in Geographical Distribution and Origin of Species.* University Press, Campbridge, 1922.

[100] J. C. Willis and G. U. Yule. Some statististics of evolution and geographical distribution in plants and animals, and their significance. *Nature*, 109:177, 1922.

[101] J.-G. Young, A. Allard, L. Hébert-Dufresne, and L. J. Dubé. Unveiling Hidden Communities Through Cascading Detection on Network Structures. *arXiv:1211.1364*, 2012.

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

[103] D. Zanette and M. Montemurro. Dynamics of text generation with realistic Zipf’s distribution. *Journal of Quantitative Linguistics*, 12:29, 2005.

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