Study of the properties and modeling of complex social graphs

Thibaud Trolliet

UNIVERSITÉ CÔTE D'AZUR

ÉCOLE DOCTORALE
SCIENCES ET TECHNOLOGIES DE L'INFORMATION ET DE LA COMMUNICATION

THÈSE DE DOCTORAT

Étude des propriétés et modélisation de graphes sociaux complexes

Thibaud TROLLIET

INRIA Sophia-Antipolis

Présentée en vue de l'obtention du grade de docteur en Informatique

de l' Université Côte d'Azur

Dirigée par : Frédéric Giroire

Soutenue le : 25 juin 2021

Devant le jury, composé de :

Augustin Chaintreau, Assistant Professor, Columbia University
Christophe Crespelle, MC, UCBL
Frédéric Giroire, DR, CNRS
Jean-Loup Guillaume, Professeur, La Rochelle Université
Clémence Magnien, DR, CNRS
Giovanni Neglia, CR, INRIA
Stéphane Pérennes, DR, INRIA
Pawel Pralat, Professor, Ryerson University

UCA JEDI INITIATIVE D'EXCELLENCE

Inria

Étude des propriétés et modélisation de graphes sociaux complexes

Jury :

Directeur de thèse :
Frédéric Giroire, DR, Centre National de Recherche Scientifique (CNRS)

Rapporteurs :
Christophe Crespelle, MC, Université Claude Bernard Lyon 1 (UCBL1)
Jean-Loup Guillaume, Professeur, Laboratoire L3I, La Rochelle Université

Examinateurs :
Augustin Chaintreau, Assistant Professor, Columbia University
Clémence Magnien, DR, Centre National de Recherche Scientifique (CNRS)
Giovanni Neglia, CR, Institut National de Recherche en Informatique et en Automatique (INRIA)
Pawel Pralat, Professor, Ryerson University

Invité :
Stéphane Pérennes, DR, Institut National de Recherche en Informatique et en Automatique (INRIA)

Résumé

L'émergence rapide des réseaux sociaux lors des deux dernières décennies, leur impact majeur sur la société actuelle, ainsi que l'accès récent à de grandes quantités de données les concernant, a amené un fort attrait pour l'étude des réseaux complexes. De nombreux modèles de graphes aléatoires ont été proposé afin de reproduire ces réseaux et leurs propriétés - faible diamètre, distribution des degrés en loi de puissance, présence de communautés, ... Cependant, du fait de la complexité de l'étude théorique de leurs propriétés, les modèles proposés sont souvent conçus pour des graphes non orientés et se concentrent sur l'émergence d'une ou deux propriétés à la fois.

Cette thèse a pour but de développer des modèles de graphes aléatoires suffisamment généraux pour reproduire de nombreuses propriétés complexes observées dans les réseaux du monde réel. En particulier, nous présentons des modèles permettant de construire des graphes avec des distributions de degrés quelconques, des graphes dirigés avec un haut coefficient de clustering, et des hypergraphes avec distribution de degrés en loi de puissance et une forte présence de communautés. Nous étudions analytiquement et expérimentalement les propriétés des modèles présentées. Nous développons enfin un outil utilisant les chaînes de Markov pour calculer la distribution des degrés de modèles d'attachement préférentiel.

Afin de s'aider dans la construction de ces modèles, nous étudions dans cette thèse deux réseaux complexes de grande taille : un graphe dirigé de l'ensemble des liens d'abonnements sur le réseau social Twitter, avec 505 millions de comptes et 23 milliards d'abonnements ; et un hypergraphe de co-publications scientifiques extraites de la base de données Scopus, contenant 2.2 millions d'auteurs distincts et 3.9 millions de publications. Des propriétés atypiques émergent de l'étude de ces deux graphes, provenant en particulier de leur caractère dirigé et d'hypergraphe, et qu'aucun modèle de la littérature ne permettait jusqu'alors de reproduire.

Mots clés : Systèmes complexes, Réseaux sociaux, Twitter, Modélisation, Graphes dirigés, Hypergraphes.

Abstract

The rapid emergence of social networks, their major impact on today's society, and the recent access to large amounts of data about them has led to a strong interest in the study of complex networks in last decades. Many random graph models have been proposed to reproduce these networks and their properties - small diameter, power-law degree distribution, presence of communities, ... However, due to the complexity of their theoretical study, the proposed models are often designed for undirected graphs and focus on the emergence of one or two properties at a time. This thesis aims at developing models of random graphs that are general enough to reproduce many complex properties observed in real-world networks. In particular, we present models for constructing graphs with arbitrary degree distributions, directed graphs with a high clustering coefficient, and hypergraphs with power-law degree distributions and a strong presence of communities. We study analytically and experimentally the properties of the presented models. We develop a tool using Markov chains to compute the degree distribution of preferential attachment models. In order to help in the construction of these models, we study in this thesis two large complex networks: a directed network with all followings on Twitter, with 505 million accounts and 23 billion followings; and a hypergraph of scientific co-publications extracted from the Scopus database, containing 2.2 million authors and 3.9 million publications. Atypical properties emerge from the study of these two graphs, coming in particular from their directed and hypergraph character, that no model of the literature could reproduce until now.

Keywords : Complex Networks, Social Networks, Twitter, Modeling, Directed graphs, Hypergraphs.

Contents

Chapter 1

Introduction

1.1 Networks and graphs

A *network* is defined as a group or system of interconnected people or things. This broad definition encompasses a lot of real-world phenomena: a network may as well represent Internet traffic, DNA interactions, food chains, financial transactions in banks, social interactions between people, Youtube recommendations, etc. To represent this set of entities and their interactions, a mathematical tool is often used: graphs. A *graph* is defined as a pair $G = (V, E)$, where V is a set of elements called *vertices* and E is a set of pairs of vertices called *edges*. The equivalence between real-world networks and the mathematical tool is immediate: the vertices correspond to the entities, while the edges correspond to the interactions between entities. For instance for financial transactions, vertices could be banks and edges money transfers between them.
While studying for several decades those different networks through graphs, a surprising phenomenon has emerged: a lot of them have common properties, despite their multidisciplinary domains. Those networks with non-trivial topological features are called *complex networks*. In opposition to lattices or fully random graphs, real-world networks exhibit some common features and are thus considered as complex networks. The most commonly observed properties are:

- **Small diameter:** Any two nodes of the network can be reached following at most $\mathcal{O}(log(N))$ links, with N the size of the network. Since the experiment of Milgram in 1967 [29], in which he studied this property in the real-world social network by sending letters by mail (Alice knows someone who knows someone who... who knows Bob), this property has been exhibited in a lot of studied real-world networks from biology [2], the web [3], and many others [46]. Actually, this property can be explained as being a consequence of the randomness of the studied networks. Indeed, a small diameter is found in random graphs built from the simplest random models, as the Ernós-Reyni model [21] in which each couple of nodes has a probability p to be connected by an edge. While lattices exhibits large diameter, randomness is sufficient to explain the small diameter property.

- **Degree distribution following a power-law:** When looking at the number of occurrences of the degrees of the nodes (where the *degree* of a node is defined as its number of connections), these *degree distributions* are often

found to be heterogeneous in real-world networks. More specifically, it is often considered as following a power-law distribution $P(i) \propto i^{-\alpha}$, with exponent α often observed between 2 and 3 - see for instance [31] and its references. This means that we observe nodes with really high degrees in comparison to others; for instance, people with really high incomes in economics, celebrities with a lot of followers on social media, ... This phenomena has various consequences: for instance on diffusion, an information shared by a celebrity will have a lot of visibility, whether this information is advertisement, fake news, prevention message, ... Also in security: for instance, if an infrastructure concentrates a lot of internet traffic, its malfunctioning could severely impact the traffic. An explanation for the emergence of this property is the "rich-get-richer" phenomena: the higher the degree of an entity is (e.g., the more famous is a celebrity or a video), the more connections it will receive (e.g. because people will talk about it, or recommendation systems will suggest it).

- **High clustering:** Real-world networks tend to form clusters, i.e., sets of entities that interact more strongly with each other than with the rest of the entities. In a graph, this amounts to a group of nodes that are highly connected to each other and weakly connected to the rest of the network. This property has been studied using two approaches:

  - The **clustering coefficient** [6, 47], which is the probability for two nodes linked by a path of length 2 to be connected. A social equivalent formulation is "what is the probability that a friend of my friend is also my friend?". In real-world social networks, this is observed to be way higher than what random graphs would give.
  - The **community detection**, which looks for an efficient way to separate nodes into communities. One of the mainly used metric is the modularity [32], which looks for an efficient partitioning of nodes into distinct communities by comparing the actual presence of a link between two nodes with the probability that this link is present in a random graph with the same degree distribution. Here again, modularity is observed as way higher in real-world complex networks than in random graphs.

  The presence of communities in real-world social networks is intuitive. However, finding the communities a posteriori with only an observation of the graph is a complicated problem. This is still a highly studied field of research due to its numerous applications: classification of nodes, link recommendations [15, 39] (e.g. in order to recommend the most suitable videos, musics or friends) or link prediction [23] (e.g. in order to estimate the future infected people by a contagious disease), etc.

These properties are intrinsic parts of real-world networks. It is therefore important to know how to properly characterize, model, and analyze those properties. In this thesis, I will focus on the two last presented properties: the power-law degree distribution and the high clustering. I will observe them in real-world networks and develop new models in order to build random graphs that have those properties.

1.2 Social networks

As described earlier, complex networks are present in many fields. In this thesis I focus on a particular type of networks: social networks. The general definition is a network of social interactions and personal relationships. This is a broad definition, which encompasses for instance Youtube followers' network, scientific collaboration network, and so on. But last decades has seen the emergence of a particular type of social networks: the *online social networks (OSNs)*. Those can be defined as a dedicated website or application which enables users to communicate with each other by posting information, comments, messages, images, etc. Facebook, Twitter, Instagram, ... are some known examples of OSNs.

While most of them have only appeared fewer than twenty years ago, they managed to take an important place in Today's society. For instance, Facebook counts 2.7 billion monthly active users in October 2020, and 1.8 billion daily active users [36]; Twitter counts 330 million monthly active users and 145 million daily active users [40]. In 2019, Internet users spent on average 2 hours and 25 minutes daily on OSNs [41]. They also have a huge impact on information: for instance, around one third of the UK and France population uses OSNs as a source of information [10, 33].

This new media has therefore managed to establish itself as an important part of society in just a few years. This leads to two important consequences: on one hand, the study of those OSNs is really interesting to enlighten social phenomena: it is the first time in history that we have access to such huge amount of social data, enabling precise statistics on social interactions. From the study of those OSNs can for instance emerges discovery on structural shapes of social interactions, or the way information spreads in a social network, ... On the other hand, the really fast growth of those OSNs makes it difficult to control the apparition of unwanted phenomena and regulate them, leading to some drawbacks we only begin to apprehend. We can mention for example:

- the presence of fake news: 86% of online global citizens believe they've been exposed to them, and 86% of those 86% admit they have believed them at least once [24]. They are suspected to have played major role in recent political events, such as Brexit or the 2016 US presidential elections [11, 38].

- the strong presence of bots: non humans accounts on social media whose purpose is often to influence opinion, either political or for marketing [7, 8].

- the apparition of a polarization: targeted recommendations are useful in order to get information we like, but have the drawback to hide the ones we are not interested to. This leads to a one-sided presentation of facts, consequence of what is called the *filter bubble* [34]. This leads to a polarization of opinions, in particular in the context of politics [27, 44].

OSNs have thus acquired a strong influence, both with advantages and drawbacks. It is therefore as interesting as essential to study them. However, practicing real experiments on OSNs may be problematic, firstly for technical reasons (they are often very large and with limited public API to access their information), secondly for ethical reasons (for instance experiments on the propagation of fake news can hardly involve real human-being who will believe those informations). Therefore, it

is important to have realistic models of OSNs, which allow experiments to be conducted on simulated networks whose properties are close to those of the real-world networks.

In this PhD, I had the opportunity to study two very large datasets with uncommon properties. The study of those complex networks enlighten new interesting properties and gave useful information in order to develop models building random graphs close to real-world complex networks. The two datasets are:

- A **Twitter followers' network**. Gabielkov and Legout [19] crawled in 2012 all the accounts of Twitter and the different followers they have. Each account is then considered as a node, and an edge is put between nodes u and v if the account associated to u follows the one associated to v. This leads to a directed graph of around 505 million nodes and 23 billion edges, making it one of the biggest directed network available nowadays. I studied the in-, out-, and bidirectionnal-degree distributions, the correlations between those in-, out-, and bidirectionnal-degrees, as well as the values of the clustering coefficients using different definitions - see Chapters 3 and 6.

- A **scientific collaboration network**. This study was conducted in partnership with economists from the laboratory of GREDEG and of the SKEMA Buisness School, with the main purpose of studying the impact of research fundings on productivity and pluridisciplinarity. They crawled data from Scopus, a database of published papers in various domains, in order to get all metadata of those papers from 1990 to 2018. From those data, I built a hypergraph of co-publications, where each node is an author and each hyperedge is a paper between authors. This leads to a hypergraph with around 2.2 million nodes and 3.9 million hyperedges. My main study of this network had been on its communities: I extracted a partition of communities and studied different properties of those communities such as their sizes, repartitions, degree distributions, ... Most of those results are gathered in Section 5.5. I propose a new hypergraph model with communities in order to reproduce the observed properties, see Chapter 5.

Note that both those datasets have some interesting specificities, in particular, namely being directed and being a hypergaph. A *directed graph* is a graph in which edges have directions, i.e., an edge is going from a node to another. A *hypergraph* is a generalisation of undirected graphs. It is defined as $H = (V, E)$ where V is a set of vertices and E a set of hyperedges, where a *hyperedge* is a set of vertices. A graph thus is a hypergraph in which any hyperedge is of size 2.

While undirected networks have received a lot of attention and had been deeply studied, there is still a lot to do for directed networks. For instance, most of the introduced metrics in order to quantify the properties discussed in Section 1.1 are suited for undirected network, such as the initial clustering coefficient definition. However, a lot of real-world networks are actually directed: Twitter, Instagram, the World Wide Web, the food chain network, ...

Likewise, many real-world networks would be better represented using hypergraphs (co-publications, chemical reactions, communication via videoconferencing tools, ...). However, studies of real-world hypergraphs are usually done by transforming the hypergraph into a graph, resulting in a loss of information during the

process. Indeed, a hyper-edge allows to represent the simultaneous interaction between a group of individuals, when an edge only represents the interaction between two given individuals. Let's take the example of co-publications. To represent a co-publication between several authors, a usual technique is to build a clique between the different authors. This transformation leads to a loss of information when these edges are "mixed" with those of other co-publications: it becomes impossible to distinguish the groups of original authors. By representing them with hyper-edges, it allows to keep the information associated with each publication.

In order to correctly model real-world networks, it is important to develop models suitable for directed and hypergraph networks, in particular with the properties discussed previously: a small diameter, a power-law degree distribution, and the presence of communities.

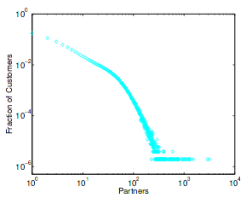
1.3 Models of random graphs

With the will to understand the real-world interactions, plenty of models of those networks have been proposed. A *model* is defined as a simple description of a system or process that can be used in calculations or predictions of what might happen. In other words, it is a simplified representation of the world through a few mathematical rules, supposed to approximate the real-world laws. Those rules are chosen such that the output of the model, a simulated network, looks as much as possible as the real-world network we want to model. The fewer rules one chooses, the simpler the model is, but also the further away it is from reality. Conversely, the more complex the model, the closer the simulated networks is to reality, but the more complicated the model is to analyze and interpret. A balance has to be found between simplicity and realism.

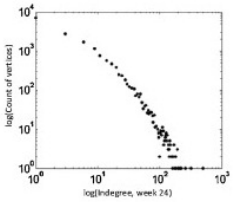
A current branch of network modeling research uses graphs. In this thesis, I focus on **probabilistic models of random growing graphs** (also called *random growth models*). In other words, we focus on models based on probabilistic rules to build random graphs whose sizes evolve over time. A lot of such models have been proposed in the literature - see Chapter 2 for a quick survey. Most of them rely on only few rules in order to stay analyzable. However, many real-world networks with various properties do not fit in the models currently proposed:

- First of all, as discussed in Section 1.2, a lot of complex networks are **not correctly represented by an undirected graph**. It is the case for networks with interactions involving more than two entities, for which hypergraphs are well-suited. Few models have focused on the creation of hypergraphs; the Avin et al. [4] model is one of them.

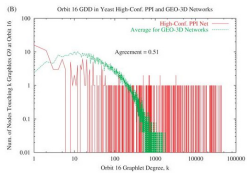
- Most of the existing models **do not recover communities**. As discussed in Section 1.1, many real-world networks have clusters, and those clusters impact the dynamic of the network, such as diffusion [17, 12]. A few models have been proposed in order to build graphs with communities [45, 25]; probably the best known is the Stochastic Block Model [22] (SBM). This model sets the number of desired communities r, then assigns each node to a random community; finally, each pair of nodes receives an edge with a probability dependent on a probability matrix P containing the probabilities of finding a link between



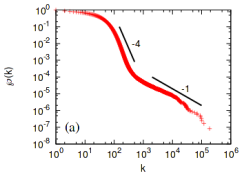
(a) Degree distribution of the number of unique callers and callees from a mobile phone operator. [37]



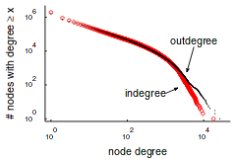
(b) In-degree distribution between shop-to-shop recommendations from an online marketplace. [42]



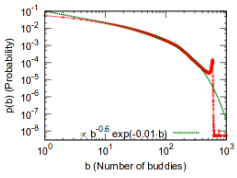
(c) Graphlet degree distribution from a biological model. [35]



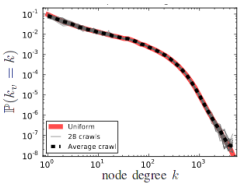
(d) Degree distribution of users of Cyworld, the largest online social network of South Korea. [1]



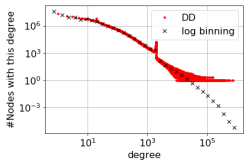
(e) Degree distributions of users of Flickr, an online social network. [14]



(f) Degree distribution of the length of the contact list in Microsoft Messenger network. [28]



(g) Degree distribution of the number of friends from FaceBook, a social network. [20]



(h) Out-degree distribution of the number of followees on Twitter. [43]

Figure 1.1: Degree distributions extracted from different seminal papers studying networks from various domains.

two given communities. However, this model has the disadvantage of being non-dynamic, and of creating a homogeneous degree distribution. More recently, the geometric preferential attachment model [18] has been shown to build graphs with both a power-law degree distribution and the presence of communities [48].

- Finally, while most models have focus on building graphs with degree distribution in power-law, a lot of real-world networks exhibit degree distributions which are not following a power-law. In [13], Broido and Clauset study the degree distribution of nearly 1000 networks from various domains, and conclude that "fewer than 36 networks (4%) exhibit the strongest level of evidence for scale-free structure". As a complement, Figure 1.1 presents degree distributions from seminal papers in various fields of the literature that clearly do not follow power-law degree distributions, to show their diversity. Models able to build such atypic distributions remained to be done.

In this thesis, I present **generalized models, able to deal with those complex properties**, and encompassing classic models from literature such as the Barabási-Albert model [5], the Chung-Lu model [16], the Bollobás et al. model [9], the Avin et al. model [4], and others. More precisely, I introduce and analyse the following models:

- A model for directed graphs with a high value of clustering coefficient of interests (*icc*, introduced in Chapter 3);
- A model able to build graphs with almost any wanted degree distribution;

- A generalized preferential attachment model for hypergraphs encompassing some classic models from literature;
- A model able to build hypergraphs with power-law degree distributions and the presence of communities;
- A model with mixed directed and bidirectional edges in order to recover some Twitter properties.

For any of those models, I show rigorously the shape of the degree distribution (power-law in all cases but the model for general degree distributions). In particular, the proof for the generalized preferential attachment model for hypergraphs gathers under a same proof the power-law degree distribution property for a whole range of classic preferential attachment models from the literature. I also **develop a new framework to study rigorously the k-dimensional degree distribution(s) of some complex models**, from classical ones to more complicated ones introduced in this thesis.

1.4 Contributions of the thesis

In summary, the contributions of this thesis are the following:

- I conduct experiments on a number of online social networks, in particular:
  - A Twitter followers' network from 2012, with 505 million nodes and 23 billion edges. The directed aspect of this network brings interesting properties, such as a high presence of bidirectional edges, and a high correlation between out- and bi-directional edges but low correlation between in/out and in/bi-directional edges.
  - A co-publication network from data extracted from Scopus between 1990 and 2018, transformed into a hypergraph with around 2.2 million nodes and 3.9 million hyperedges. I study in particular the presence of communities and their specificities.
- I propose a new clustering coefficient for directed graphs with information links, namely the *clustering coefficient of interest*, in order to capture the information part present in some OSNs. I explore it - as well as the other clustering coefficients from the previous literature - in the Twitter graph, since Twitter is known to be both a social and information media [26, 30].
- I develop a new tool to compute rigorously the degree distribution(s) of complex models, in particular those with more than one degree such as directed networks. I apply it on classical models in order to retrieve known results as well as to find unknown ones (for instance the bivariate distribution of the Bollobás et al. model). I also apply it on a new 3-dimensional Twitter's model with both directed and bidirectionnal edges.
- I introduce several new models to represent the studied networks, in particular:
  - A directed model recovering some properties of the Twitter's network (Chapter 6);

  - A directed model with a high value of interest clustering coefficient (Chapter 3);
  - A model able to build graphs with (almost) any wanted given degree distribution (Chapter 4);
  - A model for hypergraphs with power-law degree distributions and communities (Chapter 5).

  For each of those models, I compute rigorously the degree distributions.

The manuscript is organized as follow. In Chapter 2, I do a survey of the classical random models for graphs of the literature, and present the classical method used to compute the degree distribution of some of those models. Chapters 3 to 6 contain the new contributions of this thesis. In Chapter 3, I introduce the new clustering coefficient for directed interest networks, and study it on the Twitter graph. In Chapter 4, I present the new model capable to build graphs with any wanted degree distribution. In Chapter 5, I present the new model for hypergraph with communities, study its properties and compare it to the copublications hypergraph extracted from the Scopus data. Finally in Chapter 6, I present the new framework to analyse the degree distribution of complex models. I conclude in Chapter 7.

1.5 List of Publications

International conferences:

- Interest Clustering Coefficient: a New Metric for Directed Networks like Twitter - *T.Trolliet, N.Cohen, F.Giroire, L.Hogie, S.Pérennes* (Complex Networks 2020, https://hal.archives-ouvertes.fr/hal-03052083)
- A Preferential Attachment Model for any Real or Theoretical Degree Distribution - *F.Giroire, S.Pérennes, T.Trolliet* (Complex Networks 2020, https://hal.archives-ouvertes.fr/hal-03052144)

National conferences:

- Coefficient de Clustering d'intérêt: une nouvelle métrique pour les graphes dirigés comme Twitter - *T Trolliet, N Cohen, F Giroire, L Hogie, S Pérennes* (Algotel 2020, https://hal.archives-ouvertes.fr/hal-02872779)
- Revisiter l'Attachement Préférentiel, et ses applications aux Réseaux Sociaux - *G.Ducoffe, F.Giroire, S.Pérennes, T.Trolliet* (Algotel 2020, https://hal.archives-ouvertes.fr/hal-02872772)

Submitted:

- Preferential Attachment Hypergraph with High Modularity - *Frédéric Giroire, Nicolas Nisse, Thibaud Trolliet, Malgorzata Sulkowska* (Submitted to WG2021, https://hal.archives-ouvertes.fr/hal-03154836v1)
- Revisiting Preferential Attachment - *Guillaume Ducoffe, Frédéric Giroire, Stéphane Pérennes, Thibaud Trolliet*

- Interest Clustering Coefficient: a New Metric for Directed Networks like Twitter - *T.Trolliet, N.Cohen, F.Giroire, L.Hogie, S.Pérennes* (Submitted to Journal of Complex Networks, https://arxiv.org/abs/2008.00517)
- A Preferential Attachment Model for any Real or Theoretical Degree Distribution - *F.Giroire, S.Pérennes, T.Trolliet* (Submitted to Theoretical Computer Science)
- A Preferential Attachment Model for any Real or Theoretical Degree Distribution - *F.Giroire, S.Pérennes, T.Trolliet* (Submitted to Algotel 2021)
- Modèle d'attachement préférentiel pour hypergraphes avec communautés - *Frédéric Giroire, Nicolas Nisse, Thibaud Trolliet, Malgorzata Sulkowska* (Submitted to Algotel 2021)

Bibliography

[1] Yong-Yeol Ahn, Seungyeop Han, Haewoon Kwak, Sue Moon, and Hawoong Jeong. Analysis of topological characteristics of huge online social networking services. In *Proceedings of the 16th int. conference on World Wide Web*, pages 835–844, 2007.

[2] Reka Albert. Scale-free networks in cell biology. *Journal of cell science*, 118(21):4947–4957, 2005.

[3] Réka Albert, Hawoong Jeong, and Albert-László Barabási. Diameter of the world-wide web. *nature*, 401(6749):130–131, 1999.

[4] C. Avin, Z. Lotker, Y. Nahum, and D. Peleg. Random preferential attachment hypergraph. In *International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 398–405. ACM, 2019.

[5] A.L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[6] Alain Barrat and Martin Weigt. On the properties of small-world network models. *The European Physical Journal B-Condensed Matter and Complex Systems*, 13(3):547–560, 2000.

[7] Marco T Bastos and Dan Mercea. The brexit botnet and user-generated hyperpartisan news. *Social science computer review*, 37(1):38–54, 2019.

[8] Alessandro Bessi and Emilio Ferrara. Social bots distort the 2016 us presidential election online discussion. *First Monday*, 21(11-7), 2016.

[9] Béla Bollobás, Christian Borgs, Jennifer T Chayes, and Oliver Riordan. Directed scale-free graphs. In *SODA*, volume 3, pages 132–139, 2003.

[10] Noémie Bonnin. Where do people get their news? https://medium.com/oxford-university/where-do-people-get-their-news-8e850a0dea03, 2017.

[11] Alexandre Bovet and Hernán A Makse. Influence of fake news in twitter during the 2016 us presidential election. *Nature communications*, 10(1):1–14, 2019.

[12] Tom Britton, Maria Deijfen, Andreas N Lagerås, and Mathias Lindholm. Epidemics on random graphs with tunable clustering. *Journal of Applied Probability*, 45(3):743–756, 2008.

[13] Anna Broido and Aaron Clauset. Scale-free networks are rare. *Nature communications*, 2019.

[14] Meeyoung Cha, Alan Mislove, and Krishna P Gummadi. A measurement-driven analysis of information propagation in the flickr social network. In *Proceedings of the 18th international conference on World wide web*, pages 721–730, 2009.

[15] Jilin Chen, Werner Geyer, Casey Dugan, Michael Muller, and Ido Guy. Make new friends, but keep the old: recommending people on social networking sites. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 201–210, 2009.

[16] F. Chung and L. Lu. The average distances in random graphs with given expected degrees. *P. Natl. Acad. Sci. USA*, 99(25):15879–15882, 2002.

[17] Emilie Coupechoux and Marc Lelarge. Impact of clustering on diffusions and contagions in random networks. In *International Conference on NETwork Games, Control and Optimization (NetGCooP 2011)*, pages 1–7. IEEE, 2011.

[18] Abraham D Flaxman, Alan M Frieze, and Juan Vera. A geometric preferential attachment model of networks ii. *Internet Mathematics*, 4(1):87–111, 2007.

[19] Maksym Gabielkov and Arnaud Legout. The complete picture of the twitter social graph. In *Proceedings of the 2012 ACM conference on CoNEXT student workshop*, pages 19–20. ACM, 2012.

[20] Minas Gjoka, Maciej Kurant, Carter T Butts, and Athina Markopoulou. Walking in facebook: A case study of unbiased sampling of osns. In *IEEE INFOCOM*, 2010.

[21] Alexander K Hartmann and Marc Mézard. Distribution of diameters for erdős-rényi random graphs. *Physical Review E*, 97(3):032128, 2018.

[22] P. W. Holland, K. B. Laskey, and S. Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

[23] Zan Huang. Link prediction based on graph topology: The predictive value of generalized clustering coefficient. *Available at SSRN 1634014*, 2010.

[24] Ipsos. https://www.ipsos.com/en-us/news-polls/cigi-fake-news-global-epidemic.

[25] Konstantin Klemm and Victor M Eguiluz. Highly clustered scale-free networks. *Physical Review E*, 65(3):036123, 2002.

[26] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM, 2010.

[27] Jae Kook Lee, Jihyang Choi, Cheonsoo Kim, and Yonghwan Kim. Social media, network heterogeneity, and opinion polarization. *Journal of communication*, 64(4):702–722, 2014.

[28] Jure Leskovec and Eric Horvitz. Planetary-scale views on a large instant-messaging network. In *Proc. of the 17th international conference on World Wide Web*, 2008.

[29] Stanley Milgram. The small world problem. *Psychology today*, 2(1):60–67, 1967.

[30] Seth A Myers, Aneesh Sharma, Pankaj Gupta, and Jimmy Lin. Information network or social network?: the structure of the twitter follow graph. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 493–498. ACM, 2014.

[31] Mark EJ Newman. Power laws, pareto distributions and zipf's law. *Contemporary physics*, 46(5):323–351, 2005.

[32] Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.

[33] Rasmus Kleis Nielsen. Les réseaux sociaux première source d'info en ligne chez les personnes sensibles aux théories du complot. https://www.francetvinfo.fr/internet/reseaux-sociaux/info-franceinfo-les-reseaux-sociaux-premiere-source-d-info-en-ligne-chez-les-personnes-sensibles-aux-theories-du-complot_3191963.html, 2019.

[34] Eli Pariser. *The filter bubble: What the Internet is hiding from you*. Penguin UK, 2011.

[35] Nataša Pržulj. Biological network comparison using graphlet degree distribution. *Bioinformatics*, 23(2):e177–e183, 2007.

[36] Facebook Investor Relations. https://investor.fb.com/investor-news/press-release-details/2021/Facebook-Reports-Fourth-Quarter-and-Full-Year-2020-Results/default.aspx.

[37] Mukund Seshadri, Sridhar Machiraju, Ashwin Sridharan, Jean Bolot, Christos Faloutsos, and Jure Leskove. Mobile call graphs: beyond power-law and lognormal distributions. In *ACM SIGKDD*, pages 596–604, 2008.

[38] Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Alessandro Flammini, and Filippo Menczer. The spread of fake news by social bots. *arXiv preprint arXiv:1707.07592*, 96:104, 2017.

[39] Nitai B Silva, Ren Tsang, George DC Cavalcanti, and Jyh Tsang. A graph-based friend recommendation system using genetic algorithm. In *IEEE congress on evolutionary computation*, pages 1–7. IEEE, 2010.

[40] Statista. https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/.

[41] Statista. https://www.statista.com/statistics/433871/daily-social-media-usage-worldwide/.

[42] Andrew T Stephen and Olivier Toubia. Explaining the power-law degree distribution in a social commerce network. *Social Networks*, 31(4):262–270, 2009.

[43] Thibaud Trolliet, Nathann Cohen, Frédéric Giroire, Luc Hogie, and Stéphane Pérennes. Interest clustering coefficient: a new metric for directed networks like twitter. *arXiv preprint arXiv:2008.00517*, 2020.

[44] Joshua A Tucker, Andrew Guess, Pablo Barberá, Cristian Vaccari, Alexandra Siegel, Sergey Sanovich, Denis Stukal, and Brendan Nyhan. Social media, political polarization, and political disinformation: A review of the scientific literature. *Political polarization, and political disinformation: a review of the scientific literature (March 19, 2018)*, 2018.

[45] Alexei Vázquez. Growing network with local rules: Preferential attachment, clustering hierarchy, and degree correlations. *Physical Review E*, 67(5):056104, 2003.

[46] Duncan J Watts. *Small worlds: the dynamics of networks between order and randomness*. Princeton university press, 2004.

[47] Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world'networks. *nature*, 393(6684):440, 1998.

[48] Konstantin Zuev, Marián Boguná, Ginestra Bianconi, and Dmitri Krioukov. Emergence of soft communities from geometric preferential attachment. *Scientific reports*, 5(1):1–9, 2015.

Chapter 2

Preliminaries

I first present a quick survey of models from the literature which will be useful in next Chapters. Table 2.1 in Subsection 2.1.4 summarizes the properties of the presented models. I also present in Section 2.2 a classic way to compute the degree distribution of those models, using a recurrence relation on the degrees. Note that this Chapter is not intended to be an exhaustive state of the art on current models in the literature, but only to introduce some models that will help in the understanding of this thesis.

In this Chapter, a graph is defined as $G = (V, E)$ with V the set of vertices and E the set of edges. We denote by n its number of nodes, m its mean-degree, and d its diameter.

2.1 Presentation of some classic models from the literature

2.1.1 Non-growing models

Historically, the first models of random graphs introduced in the literature were **non-growing models**. Those models consider n existing nodes, and connect each pair of nodes following rules provided by the model. Unlike random growth models, there is no time evolution during the graph construction. All nodes are considered as already existing and can connect with each others.

Erdős-Réyni model

Perhaps the most famous random graphs model is the one introduced by Erdős and Réyni [14] in 1960. The Erdős-Réyni model (also often called $G_{n,p}$ model) is simple: a graph $G = (V, E)$ is built, with $|V| = n$ and such that each pair of nodes has a probability p to have an edge between them. Thus, $\mathbb{E}[|E|] = p\binom{n}{2}$, and the mean-degree is $m = pn$. This model can be extended to directed graphs simply by putting a direction to edges.

In this model, a single giant connected component exists if and only if $p > 1/n$; if $p < 1/n$, then all connected components of the graph have size $O(\log(n))$ [14]. Usually, p is chosen inversely proportional to n, so that the mean degree of the graph $m = pn$ stays independant of n. It is also known that graphs built with the $G_{n,p}$ model have small diameters, in $d = \Theta(\log(n))$ when $m > 1$ [7, 2].

Another interesting property models might want to verify is the presence of a high clustering coefficient. We remind that the global clustering coefficient (also called transitivity) is defined as ([5]):

$$\text{CC} = 3 \times \frac{\text{number of triangles in the graph}}{\text{number of paths of length 2 in the graph}}.$$

A local definition of the clustering coefficient also exists [26], see Section 3.2 for more details. In the Erdős-Réyni model, it is easy to show that the clustering coefficient is equal to p, thus inversely proportional to n and very low compared to real-world networks - often observed with a clustering coefficient higher than 0.1.

Another problem is that, since every node has a constant probability p to connect with every other node, the degree distribution follows a binomial law. This homogeneous law is far from the heterogeneous degree distributions often observed in real-world networks.

Configuration model

To overcome this, the configuration model [6, 21] proposes to put an edge between two nodes not with an uniform probability but with a probability depending of their degrees chosen beforehand. In this model, each node u receives $deg(u)$ half-edges, where $deg(u)$ is given in the model input - it corresponds to the degree distribution of the graph. Then all half-edges are matched uniformly at random. Here the degree sequence is imposed as a parameter, thereby ensuring the wanted degree distribution. As for the Erdős-Réyni model, built graphs exhibit small diameters, in $d = \Theta(\log(n))$ for power-law degree distributions [12].

The clustering coefficient in the configuration model has been studied by Newman and Park in [23]. For graphs with power-law degree distributions with exponent α, they find a dependence in $CC \sim n^{\frac{7-3\alpha}{\alpha-1}}$. The clustering coefficient thus goes to zero as n goes to the infinity when $\alpha > \frac{7}{3}$. However, surprisingly the value of the clustering coefficient diverges for power-law degree distributions with exponents less than $\frac{7}{3} \approx 2.33$ - while it is supposed to be a probability! This surprising result is due to the strong presence of loops and multiedges in the built graphs. In 2019, van der Hofstad et al. [25] computed the clustering coefficient in the *erased configuration model*, i.e. the configuration model with removal of loops and multiedges *. They find a dependence of $CC \sim n^{-(3\alpha+9)/2-2/(\alpha-1)}$. The clustering coefficient in this model always stays between 0 and 1, and goes to zero when n goes to infinity.

Watts-Stogatz model

In order to have a high clustering coefficient, Watts and Strogatz proposed in 1998 [26] a tool model ables to build graphs with both a high clustering coefficient as and a small diameter. The model starts by constructing a regular ring lattice with n nodes each connected to its K closest neighbors. Then, every half-edge is rewired with a probability β to another node chosen uniformly randomly. The choice of β changes the properties of the created graph: $\beta = 0$ leaves the regular

*I also computed this clustering coefficient at the same time as van der Hofstad et al. and arrived at the same dependence - by the time I finish writing the rigorous computation, they had published their results.

lattice untouched, leading to a high clustering coefficient but also a high diameter (of the order of n), while $\beta = 1$ gives a random graph in which the diameter is of the order of $\log(n)$ but the clustering coefficient is low. Choosing β between those extremes enables to build graphs with a small diameter and a high clustering coefficient. However, here again, while two of the main properties are satisfied, the third one is missing: indeed, each node has a degree K, thus, the degree distribution just is a Dirac function centered on K.

Stochastic Block Model

From now on, we have only considered the degree distribution, the diameter, and the clustering coefficient. In introduction, we have discussed another property often observed in real-world networks: the strong presence of communities. Quantifying this property is not easy and still is a subject of current research. Nowadays, the most used metric to quantify the presence of communities in a network is the modularity. Introduced by Newman et al. in 2004 [22], the modularity is defined as follow:

Definition 1 ([22])**.** *Let $G = (V, E)$ be a graph with at least one edge. For a partition $\mathcal{A}$ of vertices of G its modularity score on G is defined as*

$$q_{\mathcal{A}}(G) = \sum_{A \in \mathcal{A}} \left(\frac{|E(A)|}{|E|} - \left(\frac{vol(A)}{2|E|} \right)^2 \right),$$

where $E(A)$ is the set of edges within A and $vol(A) = \sum_{v \in A} deg(v)$. Modularity of G is defined as:

$$q^*(G) = \max_{\mathcal{A}} q_{\mathcal{A}}(G).$$

The first term of the sum is usually called *edge contribution*, and corresponds to the fraction of edges inside a given community A. The second term is often called *degree tax*, and corresponds to the expected fraction of edges within A that we would have in a configuration model with the same degree distribution. A high difference between those two terms means that the presence of edges inside the community is not due to randomness, but rather to a strong connection between the nodes of this community. The modularity is given by the partition of communities that maximizes $q_{\mathcal{A}}(G)$. The modularity of a graph is a value between 0 and 1: a value of 0 means that the communities are not more separated than what randomness would give, while a value close to 1 means that the communities are really well clustered.

Since we "select" the best partition of communities, even simple random graphs models have a non-null value of modularity. For instance, the modularity of Erdős-Réyni graphs (for p satisfying $1/n \leq p \leq 0.99$) is in $\Theta(\frac{1}{\sqrt{np}})$ [20]. Note that computing the theoretical modularity of random graphs is a complicated problem; really few models have an estimation of the modularity - even for the simple Erdős-Réyni model, the estimation was only done in 2020. A survey of known model modularities can be found in the appendix of [20].

Even if non-zero, the modularity of the models presented above is still lower than the one observed in real-world networks, often around 0.8. To model the presence of communities, Holland et al. introduced in 1983 the Stochastic Block Model [18]. The idea is to assign each node to a given community, then connect the nodes according

to a probability depending on their communities. The model takes as parameters the number of nodes n, a partition set of the nodes corresponding to their communities, and a probability matrix P of dimension $r \times r$, with r the number of communities. We create n nodes, each of them being assigned to a community according to the partition set given in parameters. Then, for each pair of nodes (u, v), an edge is created between u and v with probability $P_{i,j}$, where i (resp. j) is the community of u (resp. v). In the extreme case where P is diagonal, we only create edges between nodes of the same community, leading to disconnected communities.

2.1.2 Random growth model for *undirected* graphs

Barabási-Albert model[†]

The previously presented models do not take into account the dynamic evolution of the network. In 1999, Albert and Barabási [4] proposed their famous model in which the nodes arrive one after the other. The model is the following: we start at time $t = 0$ from an initial graph G_0, and we call G_t the graph at time t. At each time step t, G_t is transformed into G_{t+1} by adding a node as well as m edges attached to this node and connected to m existing nodes chosen randomly. For each edge added, an existing node has a probability to be chosen in proportion to its degree.

This model brings two novelties. Firstly, the random growth implies an evolution of the network and of node degrees. Secondly, the choice of the nodes receiving the new edges depends on the node degrees. This follows a *preferential attachment process*, i.e., a process in which nodes receive new edges depending of how many they already have. The social idea behind this is the "rich gets richer" phenomena: the more people have (money, celebrity, ...) and the easiest it is for them to get more.

An indirect effect of this preferential attachment process is the emergence of an heterogeneous degree distribution: picking nodes in proportion to their degrees implies the emergence of a degree distribution following a power-law $P(i) \propto i^{-3}$. As discussed in introduction, a lot of real-world networks exhibit degree distributions in power-laws, this explaining the success of this model.

As well as the degree distribution, this model also retrieves a small diameter, with $d = \Theta\big(\log(n)\big)$ when $m = 1$ and $d = \Theta\big(\frac{\log(n)}{\log(\log(n))}\big)$ when $m \geq 2$ [9]. However, the clustering coefficient stays low, $CC \underset{n\to+\infty}{\sim} \frac{m-1}{8}\frac{\log^2(n)}{n}$ [10].

A generalization of the Barabási-Albert model has been proposed for other preferential attachment rules. For instance, Dorogovtsev et al showed in [13] that adding an initial attractiveness, i.e., choosing nodes in proportion to $deg(u) + A$ with A a constant (instead of $deg(u)$) leads to a degree distribution following a power-law $P(i) \propto i^{-(3+A)}$. In [19], Krapivsky et al. showed that taking a non-linear preferential attachment, i.e., choosing nodes in proportion to $deg(u)^\gamma$, leads to different behavior depending of γ. For $\gamma < 1$, the degree distribution follows a stretched-exponential distribution. For $\gamma > 1$, a single node connects to almost all other

[†]Note that this model is also sometimes referred as the *preferential attachment model*. I do not stand with this definition, since a *preferential attachment process* is defined as a process in which nodes receive new edges depending of how many they already have. Thus the Barabási-Albert model is just a particular linear preferential attachment model. In the rest of thesis, I will thus refer to this model as the Barabási-Albert model.

nodes. For $\gamma = 1$, we retrieve the initial model.

Chung-Lu model

In a similar idea, Chung and Lu introduced in 2002 the following random growth model [11]: we start with an initial graph G_0. At each time step t, we add to the existing graph either (with probability p) one new vertex u and one new edge, attached to u and an existing node v chosen in proportion to its degree; or (with probability $1-p$) an edge between two existing nodes chosen in proportion to their degrees.

This model is closer to real social behavior than the Barabási-Albert model. In dynamic social networks such as OSNs, nodes continue to evolve after their arrival time; a person creates an account (arrival of a new node), and from time to time, connects to an existing node of the network (arrival of a new edge). In the Barabási-Albert model, when a new node arrives, it connects to m existing nodes; after its arrival, that node has no chance of being connected to another node that arrived before it. The Chung-Lu model allows old nodes to connect to each other. Moreover, in the Barabási-Albert model, nodes have at least degree m, since each node arrives with m edges attached to it. But m also determines the mean-degree of the graph. The mean-degree of the undirected version of the Twitter's network is around $m = 90$. So modeling this network with the Barabási-Albert model would create a graph with no nodes of degree less than 90, while in the Twitter's network, 75% of the nodes have degrees less than 90. The minimum degree in the Chung-Lu model is 1, leading to more realistic degree distributions.

Since still using a linear preferential attachment, the degree distribution follows a power-law too $P(i) \propto i^{-(2+\frac{p}{2-p})}$. I have not found any study on the distance and the clustering coefficient of this model. The similarity with the Barabási-Albert model may lead one to think that the distance will be similar. Regarding the clustering coefficient, the fact that the model add edges between old nodes might change its dependence.

Geometric Preferential Attachment Model

The last undirected model I want to present is the geometric preferential attachment model. Flaxman et al. introduced two versions of the model, one in 2006 [15] and one in 2007 [16]. I present here the second one. While nodes which receive new half-edges are still chosen in proportion to their degrees, they are also chosen depending of their proximity of the arriving node. Indeed, a geometric position is given to each node in order to represent the proximity between nodes. This can be seen as geographical proximity, but also in term of interests, ... The model is the following: we start with an initial graph G_0. At each time step t, a new node u is added to the graph, with an assigned position from the surface of the sphere in R^3 and of radius $\frac{1}{2\sqrt{\pi}}$. m edges are attached to this node and connected randomly to existing nodes such that, for every half-edge, each node v has a probability to be chosen in proportion to $deg_t(v)F(|u-v|)$, where $|u-v|$ is the distance between nodes u and v, and F is a function of that distance. Thus, nodes are chosen both depending of their degree and of their proximity to the new node.

Here again, this model build graphs with power-law degree distributions, of exponents $1+\alpha > 3$ where α is a parameter of the model. The diameter is also small,

in $O(\log(n))$. But one of the main interests to add the geometric aspect is to form clusters of nodes concentrated around their localization [16]. A few models emerged based on the same idea. For instance in [27] (with a model also called geometric preferential attachment), Zuev et al. discuss the emergence of what they call soft communities. In [1] Aiello et al. introduce the Spatial Preferential Attachment model, in which the radius of a node grows as a function of its degree: thus, the bigger the degree of the node is, the bigger is its radius, and the more chance it has to be into the proximity neighborhood of the new nodes. They show that this model follows a power-law degree distribution and has a modularity close to 1 [24]. However, the built graph is not connected: in fact, Aiello et al. expect that the majority of distinct pairs of nodes (u, v) will not have a path between them.

2.1.3 Random growth model for other types of graphs

Until now, all presented models were for undirected graphs. I present here two models built respectively for directed and hypergraph networks.

Bollobás et al. model

In 2003, Bollobás et al. proposed a model for directed graphs [8]. As in the Chung-Lu model, there are node events (addition of a node) and edge events (addition of a single edge). The model takes in parameter an initial graph G_0, three constants α, β, γ corresponding to the different event probabilities, and two constants $\delta_{in}, \delta_{out}$ corresponding to the initial attractiveness. We denote $deg_{in}(u)$ (resp. $deg_{out}(u)$) the in-degree (resp out-degree) of the node u.

We start with an initial graph G_0. At each time step t, G_t is transformed into G_{t+1} with one of these actions:

- With probability α: a new node u is added, together with an edge going from u to v where v is chosen in proportion to $deg_{in}(v) + \delta_{in}$;
- With probability β: a new edge is added, going from u to v where u and v are existing nodes chosen in proportion to $deg_{out}(u) + \delta_{out}$ and $deg_{in}(v) + \delta_{in}$ respectively;
- With probability γ: a new node w is added, together with an edge going from v to w where v is chosen in proportion to $deg_{out}(v) + \delta_{out}$.

Since the model is directed, we now have a distinction between the addition of a node with an edge leaving it (probability α), and arriving to it (probability γ) We also have a separation between the in-degree and out-degree evolutions: in this model, the choice of a node receiving (resp. leaving) an edge is dependent only to its in-degree (resp. out-degree). Hence, the evolutions of its in-degree and out-degree are independent. However, for a given node, its in-degree and out-degree are correlated, due to its age: the older a node is, the higher both those degrees will be. But those degrees are only correlated by the arrival age, whereas their evolutions are uncorrelated.

This model also add an initial attractiveness for nodes, such that the nodes are now chosen in proportion to their degrees plus an additional constant. Note that those constants are necessary due to the directed property: indeed, a node arrives

with either an in-degree of 1 and out-degree of 0, or vice versa. In both cases, one of its degree needs an initial attractiveness to have a non-zero probability to be chosen at the beginning.

This model exhibits power-law in- and out-degree distributions, of slopes $1 + \frac{1+\delta_{in}(\alpha+\gamma)}{\alpha+\beta}$ and $1+\frac{1+\delta_{out}(\alpha+\gamma)}{\gamma+\beta}$ respectively. Since the model generates directed graphs, other metrics are a priori not well-defined: to the best of my knowledge, the study of the directed diameter, clustering coefficient, and presence of communities in this model still remains to be done using generalization of undirected definitions. Among them, the study of the directed clustering coefficient is a work in progress in collaboration with Guillaume Ducoffe, Frédéric Giroire, Stéphane Pérennes and Małgorzata Sulkowska.

Avin et al. model

The last model I want to present is the one introduced by Avin et al. in 2019 [3]. While a lot of real-world networks are actually better represented with hypergraphs, really few models have focused on the building of such graphs. Most of the times, hypergraphs are transformed into graphs by replacing hyperedges with cliques.

The model proposed by Avin et al. builds hypergraphs with power-law degree distributions[‡]. It is a generalization of the Chung-Lu model, with addition of hyperedges instead of edges. We start with an initial hypergraph H_0. At each time step t, either (with probability p) we add a node and a hyperedge that contains the new node as well as $Y_t - 1$ existing nodes, or (with probability $1-p$) we add a hyperedge between Y_t nodes. Y_t is a random integer variable given as a parameter. Those Y_t (or $Y_t - 1$) nodes are chosen independently in proportion to their degrees.

Avin et al. shows that this model gives a power-law degree distribution of exponent $2 + \frac{p}{\mu-p}$, where $\mu = \mathbb{E}[Y_t]$. We notice that, in the particular case where $\forall t, Y_t = 2$, we find the exponent of the Chung-Lu model. A generalization of this model is presented in Section 5.3.1. As for the Bollobás et al. model, studying the other properties requires a transposition of the classic undirected definitions into hypergraph definitions. To the best of my knowledge, the only studied property other that the degree distribution is the one presented in Chapter 5.7 on the modularity.

2.1.4 Summary table

Table 2.1 summarizes some properties of the models discussed in previous Sections.

2.2 Classical tools for model analysis

Degree distribution being one of the major parameters of networks, it is of crucial importance to estimate its shape in models supposed to represent reality. It can be noticed from the previous Sections that the degree distribution of most presented models is well-known, and often follows a power-law. For most of them, a common method can be used to theoretically compute it. The purpose of this Section is to present the general idea of this method. I will not go into complicated computations or precise details, but only give an overview on how the method works. A rigorous

[‡]For hypergraphs, the degree of a node is defined as the number of hyperedges it belongs to.

Model	Type	RG?	DD: $P(i) = \cdots$	CC	Distance	Modularity
Erdős-Reyni	Und.	No	$\binom{k}{n-1}p^i(1-p)^{n-1-i}$	p	$\frac{\log(n)}{\log(pn)}$ (for $pn > 1$)	$\Theta(\frac{1}{\sqrt{np}})$ (for $\frac{1}{n} \leq p \leq 0.99$)
Configuration	Und.	No	anything	$n^{\frac{7-3\alpha}{\alpha-1}}$ (for PL)	$\Theta(\log(n))$	(?)
Watts-Strogatz	Und.	No	$\delta(K)$	$\frac{3(K-2)}{4(K-1)}(1-\beta)^3$	$\leq \frac{72}{\beta}\log^2(n)$	(?)
Barabási-Albert	Und.	Yes	i^{-3}	$\log(n)^2/n$	$\frac{\log(n)}{\log(\log(n))}$	$\frac{1}{m} < q^* < \frac{15}{16}$
Chung-Lu	Und.	Yes	$i^{-(2+\frac{p}{2-p})}$	(?)	(?)	(?)
Geometric PA	Und.	Yes	$i^{?}$	(?)	(?)	(?)
Spatial PA	Dir.	Yes	$i^{-(1+\frac{1}{pA_1})}$	(?)	(not connected)	$1 - o(1)$
Bollobás et al.	Dir.	Yes	$\begin{cases} P(i) & \propto i^{-(1+\frac{1+\delta_{in}(\alpha+\gamma)}{\alpha+\beta})} \\ P(o) & \propto o^{-(1+\frac{1+\delta_{out}(\alpha+\gamma)}{\gamma+\beta})} \end{cases}$	(...)	(...)	(...)
Avin et al.	Hyper.	Yes	$i^{-(2+\frac{p}{\mathbb{E}[Y_t]-p})}$	(...)	(...)	(...)

Table 2.1: Summary of some properties of the presented models. *RG* stands for *Randow Growth*, *DD* for degree distribution, and *CC* for *clustering coefficient*. *Und.* (resp. *Dir.*, *Hyper.*) stands for *undirected* (resp. *directed*, *hypergraph*). *PL* stands for *power-law*. $\delta_{i,j}$ is the Dirac function. I put a question mark when I didn't find any theoretical study of the property, and dots when the definition of the property has to be redefined.

application of it can be found in Section 5.3. We also use this method in Sections 3.6 and 4.3.1. Note also that Chapter 6 presents a new generic theoretical framework to compute rigorously degree distribution(s) of a bunch of preferential attachment models using Markov Chains. I do not detail here the related work about Markov Chain computations since it is presented in Section 6.1.1.

To illustrate the method, I apply it on the Chung-Lu model. The general concept can then be easily generalized. Let us remind the Chung-Lu model: we start with a graph G_0. At each time step t, the graph $G_t = (V_t, E_t)$ is transformed into G_{t+1} by the following event:

- With probability p: add a new node u, pick an existing node v chosen with probability $\frac{deg(v)}{\sum_{w \in V_t} deg(w)}$, and add an edge between u and v;
- With probability $1-p$: pick two existing nodes u and v chosen with probability $\frac{deg(u)}{\sum_{w \in V_t} deg(w)}$ and $\frac{deg(v)}{\sum_{w \in V_t} deg(w)}$, and add an edge between u and v.

We are looking for the probability $P(i)$ for a node to have degree i when $t \to +\infty$. Let us call $N(i,t)$ the number of nodes of degree i at time t, then $P(i) = \lim_{t \to +\infty} \mathbb{E}\left[\frac{N(i,t)}{|V_t|}\right]$.

In our example, the number of nodes at time t is a binomial concentrated around $\mathbb{E}[|V_t|] = pt$, and the number of edges is exactly $|E_t| = t$. Here the number of edges is fixed and the number of nodes is variable; depending of the studied model this may change. Having random variables for $|V_t|$ and $|E_t|$ brings some extra-steps in the computation, that are often easily overcome thanks to the concentration of those variables around their means. Here, if $|V_t|$ had been a constant - as it is the case e.g. for the Barabási-Albert model -, we could have easily taken it out of the expectation. Instead, an extra-step has to be done in order to show that $P(i) \sim \lim_{t \to +\infty} \frac{\mathbb{E}[N(i,t)]}{\mathbb{E}[|V_t|]} = \lim_{t \to +\infty} \frac{\mathbb{E}[N(i,t)]}{pt}$ - see for instance Lemma 5 for a proof of this equivalence.

We now have to compute $\mathbb{E}[N(i,t)]$. $N(i,t)$ satisfies the following recurrence relation:

$$
\begin{aligned}
N(i,t+1) - N(i,t) =& p\delta_{i,1} && (2.1)\\
&+ p\frac{i-1}{\sum_{w \in V_t} deg(w)} N(i-1,t) && (2.2)\\
&- p\frac{i}{\sum_{w \in V_t} deg(w)} N(i,t) && (2.3)\\
&+ 2(1-p)\frac{i-1}{\sum_{w \in V_t} deg(w)} N(i-1,t) && (2.4)\\
&- 2(1-p)\frac{i}{\sum_{w \in V_t} deg(w)} N(i,t), && (2.5)
\end{aligned}
$$

where $\delta_{i,j}$ is the Kronecker delta.

This equation, sometimes called **master equation**, gives a relation between $N(i,t+1)$, $N(i+1,t)$, and $N(i,t)$. It represents the modification of the number of nodes of degree i between two consecutive time steps. Indeed at each time step, we

either add a node and an edge (probability p) or an edge (probability $1-p$). Let us consider the first case. Then, for $i > 1$, the number of nodes of degree i only changes if the node chosen to be connected to the new edge is of degree $i-1$ (it goes from degree $i-1$ to degree i, increasing $N(i,t)$ by 1) or if it is of degree i (it goes from degree i to degree $i+1$, decreasing $N(i,t)$ by 1). If a node with a different degree is chosen, then $N(i,t)$ does not change. But we know that the probability for a given node of degree $i-1$ to be chosen is $\frac{i-1}{\sum_{w\in V_t} deg(w)}$. Combining all of this, the probability that we are in the first case of the model (probability p), and pick one of the $N(i-1,t)$ nodes of degree $i-1$, is $p\frac{i-1}{\sum_{w\in V_t} deg(w)}N(i-1,t)$. This explains the second term on the master equation's right-hand side.

The three following terms follow from the same reasoning: the third term is the probability for a node of degree i to be chosen in the first case of the model, the fourth term is the probability to choose a node of degree $i-1$ in the case where we add only an edge, and the fifth term is the probability to choose a node of degree i in the case where we add only an edge. The factor 2 comes from the fact that we pick two existing nodes - one for each half-edge. Finally, with probability p a new node of degree 1 is created, explaining the first term. While the terms of the master equation can change depending on the model, for most of them, we still have a relation between $N(i,t+1)$, $N(i+1,t)$ and $N(i,t)$.

In the Chung-Lu case, the master equation can be rewritten as:

$$N(i,t+1) - N(i,t) = p\delta_{i,1} + (2-p)\frac{(i-1)N(i-1,t)}{\sum_{w\in V_t} deg(w)} - (2-p)\frac{iN(i,t)}{\sum_{w\in V_t} deg(w)}. \tag{2.6}$$

This master equation, often encounter in literature, hides some approximations. Indeed, multiple draws may happen during the same step. Thus, when picking two existing nodes, the same node can be chosen two times, leading to a non-zero probability to have nodes going from degree $i-2$ to i in only one time step. In the Chung-Lu model, it might happen during the edge event (probability $1-p$), where two nodes are chosen: if we choose the same node twice, a loop will be created around this node and its degree will be increased by two in a single time step. In Barabási-Albert model, m nodes are chosen to receive the new edges; picking the same node more than once will create multiedges and increase its degree by more than one. The same phenomena occurs in many models. However those cases are rare and can be neglected in most of the cases. Although it is often intuitively clear that those terms do not impact the result, proving it rigorously can be really challenging. Most of the proofs for models of the literature do not talk about those additional terms, or just qualitatively discuss their lack of influence without rigorously proving it. An example of how to deal with those additional terms is given in Section 5.3.2, especially Equation 5.22.

Before taking the expectation on both sides of the equation, let us discuss about the denominator term $\sum_{w\in V_t} deg(w)$. In the case of the Chung-Lu model, since $|E_t| = t$, we can directly express the sum as

$$\sum_{w\in V_t} deg(w) = 2|E_t| = 2t.$$

In this case, the denominator is not a random variable. This implies that, when

taking the expectation on both size of the master equation, it can be taken out of the expectation. However in more general cases, two complications may appear:

- The number of edges is a random variable: thus, we have to show that it is concentrated enough to have $\mathbb{E}\left[\frac{iN(i,t)}{\sum_{w\in V} deg(w)}\right] \sim \frac{\mathbb{E}[iN(i,t)]}{\sum_{w\in V} deg(w)}$.
- The attachment function $f(i)$ is not linear: then, the sum cannot be expressed as a function of the mean-degree, thus, does not have a simple expression. This problem appears in Chapter 4 where we use general attachment functions.

In both cases, those complications can be resolved thanks to what is called concentration inequalities. For instance, Hoeffding's inequality [17] states the following:

Lemma 1 (Hoeffding's inequality, [17]). *Let $Z_1, Z_2, \ldots, Z_t$ be independent random variables such that $\mathbb{P}[Z_i \in [a_i, b_i]] = 1$. Let $\delta > 0$ and $Z = \sum_{i=1}^{t} Z_i$. Then*

$$\mathbb{P}[|Z - \mathbb{E}[Z]| \geqslant \delta] \leqslant 2\exp\left\{-\frac{2\delta^2}{\sum_{i=1}^{t}(a_i - b_i)^2}\right\}.$$

This Lemma can be applied with Z_i representing the number of half-edges added at each time step. In most cases, this number is bounded and it is easy to find a couple (a, b) such that $\forall i, \mathbb{P}[Z_i \in [a, b]] = 1$. Consider for instance the Barabási-Albert model modified such that, at each time step, the number of added edges is a random variable between 1 and m (instead of exactly m). In this model, $|E_t|$ is a random variable. However, we know that at most $2m$ half-edges are added at each time step. $|E_t|$ can be express as $\sum_{w\in V} deg(w) = \sum_{i=1}^{t} Z_i$, with Z_i the number of half-edges added at time step t. Applying Lemma 1 with $Z = \sum_{w\in V} deg(w)$, $\delta = (m-1)\sqrt{2t\log(t)}$, and $\forall i$ $a_i = 1$ and $b_i = m$, gives:

$$\mathbb{P}[|Z - \mathbb{E}[Z]| \geqslant \delta] \leqslant 2\exp\left\{-\frac{4(m-1)^2 t\log(t)}{(m-1)^2 t}\right\} = \frac{2}{t^4}. \tag{2.7}$$

Thus, Z is highly concentrated around $\mathbb{E}[Z]$.

Let us come back to solving the master equation. In the Chung-Lu model case, taking the expectation on both sides give:

$$\begin{aligned}\mathbb{E}[N(i,t+1)] - \mathbb{E}[N(i,t)] = p\delta_{i,1} + (2-p)\frac{(i-1)\mathbb{E}[N(i-1,t)]}{2t} \\ - (2-p)\frac{i\mathbb{E}[N(i,t)]}{2t}.\end{aligned} \tag{2.8}$$

This equation is then solved using the following lemma:

Lemma 2 ([11], Chapter 3.3). *Let $\{a_t\}$ be a sequence satisfying the recursive relation*

$$a_{t+1} = \left(1 - \frac{b_t}{t}\right) a_t + c_t$$

where $\lim_{t\to\infty} b_t = b > 0$ and $\lim_{t\to\infty} c_t = c$. Then, the limit $\lim_{t\to\infty} \frac{a_t}{t}$ exists and

$$\lim_{t\to\infty} \frac{a_t}{t} = \frac{c}{1+b}.$$

For the Chung-Lu model, for $i > 1$, Lemma 2 can be applied with $a_t = \mathbb{E}[N(i,t)]$, $b_t = \frac{2-p}{2} \times i$, and $c_t = \frac{2-p}{2}(i-1)\frac{\mathbb{E}[N(i-1,t)]}{t}$ to obtain:

$$\lim_{t\to\infty} \frac{\mathbb{E}[N(i,t)]}{t} = pP(i) = \frac{\lim_{t\to\infty} \frac{2-p}{2}(i-1)\frac{\mathbb{E}[N(i-1,t)]}{t}}{1+\frac{2-p}{2}i} \tag{2.9}$$

$$= \frac{i-1}{\frac{2}{2-p}+i} pP(i-1) \tag{2.10}$$

Lemma 2 can also be used for $i = 1$ with $a_t = \mathbb{E}[N(1,t)]$, $b_t = \frac{2-p}{2}$, and $c_t = p$ to obtain:

$$pP(1) = \frac{p}{1+\frac{2-p}{2}}. \tag{2.11}$$

We now have a recurrence between $P(i)$ and $P(i-1)$. Iterating over $P(i)$, we finally get:

$$P(i) = P(1)\prod_{k=1}^{i-1} \frac{k}{k+1+\frac{2}{2-p}} = \Gamma(1+\frac{2}{2-p}) \times \frac{\Gamma(i)}{\Gamma(i+1+\frac{2}{2-p})}. \tag{2.12}$$

We have expressed the degree distribution of the Chung-Lu model. For i high enough, this fraction of Gamma functions can be approximated as a power-law of exponent $1+\frac{2}{2-p}$ [§]:

$$P(i) \underset{i\to+\infty}{\sim} \Gamma(1+\frac{2}{2-p}) \times i^{-(1+\frac{2}{2-p})}. \tag{2.13}$$

Taking other models with linear preferential attachments give similar results, but with different exponents. If we go back to Equation 2.6 to keep track of the exponent, we see that it actually comes from the terms $(p+2(1-p)) \times \frac{1}{\sum_{w\in V_t} deg(w)}$. Those terms differ for different models, but the same method can still be used in a similar way.

Remark 1. *The main idea of Chapter 4 is to reverse the presented master equation. Indeed, if we replace the linear attachment function by a more general function $f(i)$, we see that we also have a recurrence relation between $f(i)$ and $f(i-1)$. We have computed here the degree distribution P knowing the linear attachment function i; but we can also suppose that we know the degree distribution, and solve the master equation to compute an unknown attachment function f - see Chapter 4 for more details about this inversion and its applications.*

[§]Since $1+\frac{2}{2-p} = 2+\frac{p}{2-p}$, this result coincides with the one of Table 2.1

Bibliography

[1] William Aiello, Anthony Bonato, Colin Cooper, Jeanette Janssen, and Paweł Prałat. A spatial web graph model with local influence regions. *Internet Mathematics*, 5(1-2):175–196, 2008.

[2] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47, 2002.

[3] C. Avin, Z. Lotker, and D. Peleg. Random preferential attachment hypergraphs. *CoRR*, abs/1502.02401, 2015.

[4] A.L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[5] Alain Barrat and Martin Weigt. On the properties of small-world network models. *The European Physical Journal B-Condensed Matter and Complex Systems*, 13(3):547–560, 2000.

[6] Edward A Bender and E Rodney Canfield. The asymptotic number of labeled graphs with given degree sequences. *Journal of Combinatorial Theory, Series A*, 24(3):296–307, 1978.

[7] Béla Bollobás. The diameter of random graphs. *Transactions of the American Mathematical Society*, 267(1):41–52, 1981.

[8] Béla Bollobás, Christian Borgs, Jennifer T Chayes, and Oliver Riordan. Directed scale-free graphs. In *SODA*, volume 3, pages 132–139, 2003.

[9] Béla Bollobás and Oliver Riordan. The diameter of a scale-free random graph. *Combinatorica*, 24(1):5–34, 2004.

[10] Béla Bollobás and Oliver M Riordan. Mathematical results on scale-free random graphs. *Handbook of graphs and networks: from the genome to the internet*, pages 1–34, 2003.

[11] F. Chung and L. Lu. *Complex Graphs and Networks*. American Mathematical Society, 2006.

[12] Fan Chung and Linyuan Lu. The average distances in random graphs with given expected degrees. *Proceedings of the National Academy of Sciences*, 99(25):15879–15882, 2002.

[13] SN Dorogovtsev, JFF Mendes, and AN Samukhin. Generic scale of the" scale-free" growing networks. *arXiv preprint cond-mat/0011115*, 2000.

[14] Paul Erdős and Alfréd Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5(1):17–60, 1960.

[15] Abraham D Flaxman, Alan M Frieze, and Juan Vera. A geometric preferential attachment model of networks. *Internet Mathematics*, 3(2):187–205, 2006.

[16] Abraham D Flaxman, Alan M Frieze, and Juan Vera. A geometric preferential attachment model of networks ii. *Internet Mathematics*, 4(1):87–111, 2007.

[17] W. Hoeffding. Probability inequalities for sums of bounded random variables. *J. Am. Stat. Assoc.*, 58(301), 1963.

[18] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

[19] Paul L Krapivsky, Sidney Redner, and Francois Leyvraz. Connectivity of growing random networks. *Physical review letters*, 85(21):4629, 2000.

[20] Colin McDiarmid and Fiona Skerman. Modularity of erdős-rényi random graphs. *Random Structures & Algorithms*, 57(1):211–243, 2020.

[21] Michael Molloy and Bruce Reed. A critical point for random graphs with a given degree sequence. *Random structures & algorithms*, 6(2-3):161–180, 1995.

[22] Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.

[23] Mark EJ Newman and Juyong Park. Why social networks are different from other types of networks. *Physical review E*, 68(3):036122, 2003.

[24] Liudmila Ostroumova Prokhorenkova, Paweł Prałat, and Andrei Raigorodskii. Modularity of complex networks models. In *International Workshop on Algorithms and Models for the Web-Graph*, pages 115–126. Springer, 2016.

[25] Remco van der Hofstad, Pim Van der Hoorn, Nelly Litvak, and Clara Stegehuis. Limit theorems for assortativity and clustering in null models for scale-free networks. *arXiv preprint arXiv:1712.08097*, 2017.

[26] Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world'networks. *nature*, 393(6684):440–442, 1998.

[27] Konstantin Zuev, Marián Boguná, Ginestra Bianconi, and Dmitri Krioukov. Emergence of soft communities from geometric preferential attachment. *Scientific reports*, 5(1):1–9, 2015.

Chapter 3

Interest Clustering Coefficient: a New Metric for Directed Networks like Twitter

3.1 Introduction

Networks appear in a large number of complex systems, whether they are social, biological, economical or technological. Examples include neuronal networks, the Internet, financial transactions, online social networks, ... Most "real-world" networks have some properties that are not due to chance and that are really different from random networks or regular lattices. In this Chapter, we focus on the study of the clustering coefficient of social networks. Nodes in a network tend to form highly connected neighborhoods. This tendency can be measured by the clustering coefficient. It is classically defined for undirected networks as three times the number of triangles divided by the number of open triangles (formed by two incident edges). This clustering coefficient has been computed in many social networks and had been observed as much higher than what randomness would give. Triangles thus are of crucial interest to understand "real world" networks.

However, a large quantity of those networks are in fact directed (e.g. the web, online social networks like Instagram, financial transactions). It is for instance the case of Twitter, one of the largest and most influential social networks with 126 million daily active users [34]. In Twitter, a person can follow someone she is interested in; the resulting graph, where there is a link $u \to v$ if the account associated to the node u followed the account associated to the node v, is thus directed. In this study, we used as main dataset the snapshot of Twitter (TS in short) extracted by Gabielkov et al. as explained in [14] and made available by the authors. The TS has around 505 million nodes and 23 billion arcs, making it one of the biggest snapshots of a social network available today.

The classic definition of the clustering coefficient cannot be directly applied on directed graphs. This is why most of the studies computed it on the so-called *mutual graph*, as defined by Myers & al. in [27], i.e., on the subgraph built with only the bidirectional links. We call *mutual clustering coefficient (mcc for short)* the clustering coefficient associated with this graph. We computed this coefficient in the TS, using both exact and approximated methods. We find a value for the mcc of 10,7%. This is a high value, of the same order than the ones found in other

web social networks [26, 37].

However, this classical way to operate *leaves out 2/3 of the graph!* Indeed, the bidirectional edges only represents 35% of the edges of the TS. A way to avoid it is to consider all links as undirected and to compute the clustering coefficient of the obtained undirected graph. We call *undirected clustering coefficient (ucc for short)* the corresponding computed coefficient. Such a computation in the TS gives a value of ucc of only 0.11%. This is way lower than what was found in most undirected social networks. It is thus a necessity to introduce specific clustering coefficients for the directed graphs. More generally, when analyzing any directed datasets, it is of crucial importance to take into account the information contained in its directed part in the most adequate way.

A first way to do that is to look at the different ways to form triangles with directed edges. Fagiolo computed the expected values of clustering coefficients considering directed triangles for random graphs in [11] and illustrated his method on empirical data on world-trade flows. There are two possible orientations of triangles: transitive and cyclic triangles, see Figures 3.1b and 3.1c. Each type of triangles corresponds to a directed clustering coefficient :

- the ***transitive clustering coefficient*** (***tcc*** in short), defined as:

$$tcc = \frac{\#\text{ transitive triangles}}{\#\text{ open transitive triangles}},$$

- the ***cyclic clustering coefficient*** (***ccc*** in short), defined as:

$$ccc = \frac{3 \cdot \#\text{ cyclic triangles}}{\#\text{ open transitive triangles}}.$$

We computed both coefficients for the snapshot, obtaining $tcc = 1.9\%$ and $ccc = 1.7\%$. However, note that a large part of the transitive and cyclic triangles comes from bidirectional triangles. When removing them, we arrive to values of $tcc = 0.51\%$ and $ccc = 0.24\%$.

We believe those metrics miss an essential aspect of the Twitter graph: while the clustering coefficient were defined to represent the social cliques between people, it is not adequate to capture the information aspect of Twitter, known to be both a social and information media [18, 27]. In this work, we go one step further in the way directed relationships are modeled. We argue that in directed networks, *the best way to define a relation or similarity between two individuals (Bob and Alice) is not always by a direct link, but by a common interest*, that is, two links towards the same node (e.g., Bob $\rightarrow$ Carol and Alice $\rightarrow$ Carol). Indeed, when discussing interests, consider two nodes having similar interests. Apart from being friends, these two nodes do not have any reason to be directly connected. However, they would tend to be connected to the same out-neighbors. We exploit this to study a new notion of connections in directed networks and the new naturally associated clustering coefficient, which we name *interest clustering coefficient, or* `icc` in short, and define as follows:

$$\texttt{icc} = \frac{4 \cdot \#\text{ K22s}}{\#\text{ open K22s}},$$

where a K22* is defined as a set of four nodes in which two of them follow the two others, and an open K22 is a K22 with a missing link, see Figure 3.1d. We computed the `icc` on the Twitter snapshot, obtaining `icc` = 3.6% (3.1% when removing the bidirectional structures). This value, an order of magnitude higher than the previous clustering coefficients computed on the non bidirectional directed graph, confirm the interest of this metric. If the clustering coefficient of triangles are good metrics to capture the social aspect of a graph, the interest clustering coefficient is a good metric to capture the informational aspect.

In summary, our contributions are the following:

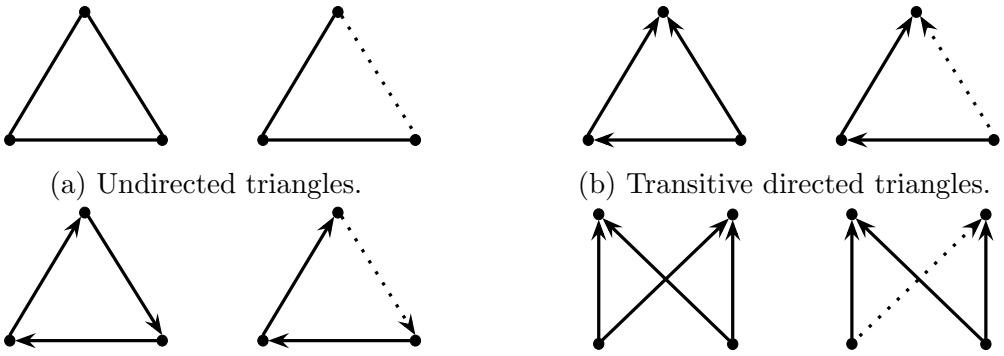
- We define a new clustering coefficient for graphs with interest links.
- We succeeded in computing it, both exactly and using sampling methods, for a snapshot of Twitter with 505 million nodes and 23 billion edges.
- We additionally provide the values of the directed and undirected clustering coefficients previously defined in the literature. We believe this is the first time that such coefficients are computed exactly for a large *directed* online social network.
- We compute this new metric as much as the previous ones on other directed datasets to highlight the differences and interests of the different metrics.
- We then propose a new random graph model to obtain random directed graphs with a high interest clustering coefficient. We prove this model follows power-law in- and out-degree distributions, and analyse the interest clustering coefficient value by simulation.
- Lastly, we discuss the usage of this new metric for link recommendation. The principle is to recommend links closing a large number of K22s (instead, classically, of triangles). We discuss the strengths/weaknesses of this method for a set of Twitter users.

The Chapter is organized as follows. We first discuss related work in Section 3.2. In Section 3.3, we present the algorithms we used to compute the values of the interest clustering coefficient, both exactly and by sampling. We discuss the results on the clustering coefficients of Twitter in Section 3.4, and of other directed datasets in Section 3.5. In Section 3.6, we propose and study a preferential attachment model providing a high interest clustering coefficient. Lastly, we discuss the use of interest clustering coefficient for link recommendation in Section 3.7.

3.2 Related Work

Complex networks. Even if the study of complex networks is an old field [36], it keeps receiving a lot of attention from the research community. The reason for this is twofold. First, a great number of very large practical systems emerged recently can be seen as complex networks, in particular online social media networks, see [24] for a

*The name comes from Graph Theory. A $K_{m,n}$ is a complete bipartite graph $G = (V_1 \cup V_2, E)$ with partitions of sizes $|V_1| = m$ and $|V_2| = n$. We consider in this Chapter a directed version of a $K_{2,2}$.



(a) Undirected triangles. (b) Transitive directed triangles.

(c) Cyclic directed triangles. (d) K22s.

Figure 3.1: Closed (left) and open (right) undirected and directed triangles and K22s.

survey. Second, with the development of big data analysis, entrepreneurs, analysts or researchers have new tools to study those huge amounts of data. Complex networks often share common properties, like small diameter [1], small average distance [40, 3, 21], heavy tail degree distributions [8, 21], high clustering [40], communities [37], etc.

Clustering coefficient. Among those properties, the clustering coefficient shows that, when two people know each other, there is a high probability that those people have common friends. The clustering coefficient has numerous important applications, such as spam detection [6], link recommendation [35, 7], information spread [15], study of biased network samples [28], performance of some neural networks [16] , etc. There are different definitions of the clustering coefficient. The *local clustering coefficient* of a node i, first introduced by Watts and Strogatz [40], is defined as the probability that two neighbors of i are also connected together. This probability can be computed as

$$CC(i) = \frac{\#\text{ triangles with the node i}}{\#\text{ connected triplets centered on i}},$$

where (# connected triplets centered on i) = $\binom{\deg(\text{i})}{2}$. From here can be defined for the whole graph a clustering coefficient as the mean of the local clustering coefficients over all the nodes of the graph:

$$CC_{g1} = \frac{1}{n}\sum_{i \in V} CC(i)$$

Another definition was first introduced by Barrat and Weigt in [4], and is called the *global clustering coefficient*, or *transitivity*. It is defined as

$$CC_g = 3 \times \frac{\#\text{ triangles in the graph}}{\#\text{ connected triplets of vertices in the graph}}.$$

We use the global clustering coefficient in this Chapter. The clustering coefficient has also been defined for weighted graphs [32, 29].

Computations for social graphs. The undirected clustering coefficient of some social networks has been provided in the literature. It has been computed on very large snapshots for Facebook [37], Microsoft Messenger [21], Flickr, and YouTube [26]. The local clustering coefficient has also been studied in the undirected mutual graph of Twitter [27]. We can also cite the values given by the Network Repository project [30], providing a large comprehensive collection of network graph data available for which it lists some basic properties. The undirected clustering coefficient is usually much higher in social networks than in random models.

Directed graphs. All these studies only consider the undirected clustering coefficient, even for directed graphs like Twitter. Fagiolo introduced definitions of directed clustering coefficients, that we named tcc and ccc [11], but those definitions had never been computed and discussed on large datasets to our knowledge, as we do in this Chapter. Moreover, we believe that these metrics are *not the most relevant ones for directed graphs with interest links.*

Computing substructures. Researchers studied methods to efficiently compute the number of triangles in a graph, as naive methods are computationally very expensive on large graphs. Two families of methods have been proposed: triangle exact counting or enumeration and estimations. In the first family, the fastest algorithm is due to Alon, Yuster, and Zwick [2] and runs in $O(m^{\frac{2\omega}{\omega+1}})$, with m the number of edges and ω the best known exponent for the fast matrix multiplication. Its current value is 2.3728, due to an algorithm of [9] improved by [20], giving a complexity of $O(m^{1.41})$ for the AYZ algorithm. However, methods using matrix multiplication cannot be used for large graphs because of their memory requirements. In practice, enumeration methods are often used, see e.g., [19, 33]. A large number of methods for approximate counting were proposed, see for example [17] and its references. The authors obtain a running time of $O(m + \frac{m^{3/2}\log n}{t\varepsilon^2})$ and a $(1 \pm \varepsilon)$ approximation. Methods to count rectangles and butterfly structures in undirected bipartite networks were also proposed in [39] and in [31]. In this Chapter, we propose an efficient enumeration algorithm to count the number of K22s and open K22s in a very large graph. We focused on the case in which only one adjacency can be stored, as this was our case for the TS. To the best of our knowledge, we are the first to consider this setting.

3.3 Computing Clustering Coefficients in Twitter

We computed the interest clustering coefficient and the triangle clustering coefficients on a directed Twitter snapshot (TS in short) that we use as a typical example of a directed social network with interest links. We used two different methods: an exact count and an estimation using sampling techniques, either with a Monte Carlo algorithm or with a sampling of the graph.

3.3.1 The Twitter Snapshot

In order to compute the different clustering coefficients of a real graph, the authors of [13] gave us access to a snapshot of the graph of the followings of Twitter. The

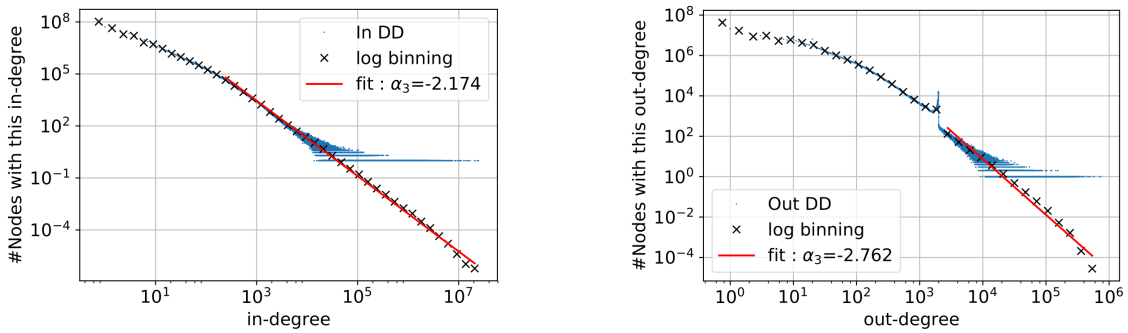


Figure 3.2: In- (Left) and out-degree (Right) distributions of the Twitter Snapshot. The obtained distribution is given by the blue points; the black crosses represent the logarithmic binning of the distribution (a mean of a given amount of points on a logarithmic scale). The red straight line is the fit of the logarithmic binning; it has slopes of -2.174 and -2.762 for the in and out degree distribution.

snapshot was collected between March 2012 and July 2012. With $n = 505$ million nodes and $m = 24$ billion links collected, this graph is the largest directed social network graph available today, to the best of our knowledge. Each node of the graph represents an account of Twitter, and there is a link between two nodes u and v, if the account u follows the account v. All account IDs have been anonymized. The snapshot is a perfect case study as Twitter is a directed social network used both as a social and an information network [27, 18]. It allows to study directed/undirected social/interest clustering coefficients.

Degree distributions of the Twitter Snapshot. We provide in Figure 3.2 the degree distributions of the TS. We fitted their tails to power law distributions. We obtained $P^-(i) = C^- i^{-2.17}$ and $P^+(i) = C^+ i^{-2.76}$, with $P^-(i)$ (respectively $P^+(i)$) the probability that a node has in-degree (resp. out-degree) i. In the following, we use the obtained values to compute the practical complexity of the algorithms.

Other references of the literature have also provided a power law fit for both distributions, see e.g., [27]. In this work, the authors obtained exponents of values 1.35 and 1.28. However, we believe that the authors did a fit on the complete distributions and not on their tails, leading to power law exponents below 2. This is why we preferred to only fit the tail. Another point of discussion would be to decide if the out-degree distribution really behaves as a power law. However, the best fit of the distributions is out of the scope of this Chapter. We just used the values provided by our fit as a possible model of the graph, but others exist.

3.3.2 Exact Count

We computed the exact numbers of K22s and open K22s in the Twitter Snapshot. Recall that we are discussing a dataset with hundreds of million nodes and billions of arcs. Results are reported in Table 3.1 and discussed in Section 3.4. We also retrieved the number of directed and undirected triangles of TS. We first discuss the complexity of algorithms for exact counting on very large graphs. We then present the algorithms we use and discuss the results.

In the rest of this Chapter, we call *top vertices* (resp. *bottom vertices*) of a K22 the vertices which are destinations (resp. sources) of the K22 edges. We call a *fork* a set of two edges of a K22 connected to the same vertex. We say that a *fork has top*

(or bottom) vertex x if both edges are connected to x and x is a top (resp. bottom) vertex of the K22. The same terminology applies to open K22s.

Trivial algorithm. The trivial algorithm would consider all quadruplets of vertices with 2 upper vertices. Then, for each quadruplet, it would check the existence of a K22 and of open K22s. There are $\binom{4}{2}\binom{n}{4}$ such quadruplets. It thus gives a complexity of $O(n^4)$. This method can thus not be considered for the TS as it would perform 6.4×10^{22} iterations.

Improved algorithm. The practical complexity can be greatly improved by only considering *connected quadruplets*, and by mutualizing the computations of the common neighbors of the in-neighbors of a vertex, as explained below. The pseudo-code is given in Algorithm 1.
The algorithm's main loop iterates on the vertices of the graph. For each vertex x, we consider its in-neighborhood $N^-(x)$. We then compute how many times a vertex w (with $w < x$ to avoid counting a K22 twice) appears in the out-neighborhoods of the vertices of $N^-(x)$. We denote it $\#occ(w)$. We use a hash table to store the value of $\#occ(w)$ in order to be able to do a single pass on each out-neighbor.
For a vertex w, any pair of its $\#occ(w)$ in-neighbors common with x forms a K22 with x and w as bottom vertices. There are hence $\binom{k}{2}$ K22s with x and w as bottom vertices. The number of K22s with x as a top vertex is then

$$\#K22(x) = \sum_{w|\#occ(w)\geq 2} \binom{\#occ(w)}{2}.$$

The number of open K22s with x as the top vertex is computed by noticing that, for any pair of vertices u and v of $N^-(x)$, we have $d^+(u) - 1 + d^+(v) - 1 - \mathbb{1}_{v\in N^+(u)} - \mathbb{1}_{u\in N^+(v)}$ open K22s containing this fork (ux, vx). We can count the number of open K22s with x as a top vertex, u as the bottom vertex of out-degree 2 (and thus another vertex v as the bottom vertex of out-degree 1). A vertex $u \in N^-(x)$ is thus in $(d^+(u)-1\sum_{v\in N^-(x)\setminus\{u\}} 1_{v\in N^+(u)})(d^-(x)-1)$ such open K22s. The only subtlety is that we count the number of arcs, which are between two vertices of $N^-(x)$, during the loop on the out-neighborhoods of the vertices of $N^-(x)$. We note this number #internalArcs. We then have:

$$\#openK22(x) = \left(\sum_{u\in N^-(x)} (d^+(u) - 1)(d^-(x) - 1)\right) - \#internalArcs.$$

Lastly, the global number of K22s (resp. open K22s) in the digraph is just the sum of the number of K22s (resp. open K22s) with a vertex x as a top vertex, as, since we only consider K22s formed with a vertex w such that $x < w$, we only count each K22 once.

Complexity of the used algorithm. The complexity thus is $m+\sum_u d^+(u)(d^+(u)-1)$. Indeed, each edge is only considered once as an in-arc and d^+-1 times as an out-arc. Note that, in the Twitter Snapshot, the sum of the squares of the degrees is equal to $8 \cdot 10^{13}$. The order of the number of iterations needed to compute the number of K22s was thus massively decreased from the 6.4×10^{33} iterations of the trivial algorithm.

Algorithm 1 Enumeration of K22s and open K22s

```
▷
Input: Digraph(V, A)
#occ=0                                                    ▷ hash table
for x ∈ V do
    #internalArcs ← 0    ▷ We count the number of arcs internal to N⁻(x) as these arcs do not form open K22s
    for v ∈ N⁻(x) do
        #openK22s += (d⁺(v) − 1)(d⁻(x) − 1)
        for w ∈ N⁺(v) \ {x} do
            #occ[w] += 1
            if w ∈ N⁻(x) then        ▷ We use a second hash table to test that.
                #internalArcs += 1
    for w with #occ[w] ≥ 2 do
        #k22 += (#occ[w] choose 2)
    #openK22s −= #internalArcs
    #occ ← 0                                              ▷ Done with a double loop
icc ← 4#K22 / #openK22
```

Complexity on graphs following a power-law degree distribution. The complexity of the algorithm on a graph built with preferential attachment can be computed as follows. We consider without loss of generality that the sum of the square of the degrees is minimum for the out-degrees (and not the in-degrees). The maximum degree is $d^+_{\max} = O(n^{1/(\alpha^+-1)})$, with α^+ the exponent of the out-degree power law distribution. Thus, the sum of the squares of the degrees, when $2 \leq \alpha^+ < 3$, is $\sum_{v\in V}(d^+(v))^2 = C^+ n \sum_{i=1}^{d^+_{\max}} \frac{i^2}{i^{\alpha^+}} \underset{n\to\infty}{\sim} C^+ n \int_{i=1}^{d^+_{\max}} \frac{1}{i^{\alpha^+-2}} = \left[\frac{C^+ n}{(3-\alpha^+)i^{\alpha^+-3}}\right]_1^{d^+_{\max}} \simeq \frac{C^+ n}{(3-\alpha^+){d^+_{\max}}^{\alpha^+-3}} = \frac{C^+}{(3-\alpha^+)} n^{1+\frac{3-\alpha^+}{\alpha^+-1}}$, where $C^+ = \frac{1}{\sum_{i\in\mathbb{N}^+} i^{\alpha^+}}$. The complexity is thus in $O(m + n^{1+\frac{3-\alpha^+}{\alpha^+-1}})$. For preferential attachment graphs with exponents between 2 and 3, this gives a complexity between $O(m+n)$ and $O(n^2)$, to be compared to the one of the naive method $O(n^4)$.

Counting the number of triangles. The number of transitive triangles can easily be computed for free while counting the K22s. When iterating over the vertices of the TS and considering the vertex x in Algorithm 1, the number `internal_arcs` of arcs between vertices of $N^-(x)$ corresponds to the number of transitive triangles for which x is the top vertex. The number of open transitive triangles with x as the top vertex is simply $d^-(x) \cdot d^+(x)$. The total number of open transitive triangles is then just the sum of this quantity over all x. The number of cyclic triangles for x can also be easily computed by counting the number of arcs from $N^+(x)$ to $N^-(x)$. Each cyclic triangle is counted three times. The number of open cyclic triangles is the same as the number of transitive triangles. We can compute the number of undirected triangles with similar methods (either on the full (but undirected) graph or on the mutual graph).

Note that the fastest methods to compute triangles in graphs have a complexity of $O(m^{1.41})$, where m is the number of edges [2]. These methods rely on fast matrix multiplications and cannot be applied for large graphs as they need to have the full matrix in memory. Moreover, our algorithms would be faster in practice for large complex networks as they are sparse graphs. The average indegree (or outdegree) has a low value of 45.6 [14] in Twitter. The complexity of the matrix methods would be of the order of $3.2 \cdot 10^{14}$ for the TS as $m = 2.3 \cdot 10^{10}$. This is higher than the practical complexity of computing the exact number of K22s (which is itself higher than the complexity of computing triangles). We discuss the obtained results with the exact count in Section 3.4.

3.3.3 Approximate Counts

As discussed later in Section 3.4, the exact count of the number of K22s and open K22s in Twitter implies massive computations. This number can be estimated using Monte Carlo Method and/or computations on a sample of the graph. We discuss both methods below. One of our goals was to see how good computations made in the literature using smaller Twitter snapshots were.

Exact `icc` on Twitter Samples.

We built samples of the TS to estimate the interest clustering coefficient. Several choices can be made to build the samples. To avoid missing nodes of high degrees (which would lead to a high variance), we sampled the arcs (and not the nodes). Given a sampling probability p, we keep an arc in the sample with probability p. We generated samples of different sizes corresponding to sampling probabilities from $p = 1/100$ to $p = 1/16000$.

Estimator of the number of K22 and open K22s. Let us call $\mathcal{A}$ the set of occurrences of a specific pattern (in our case, either a K22 or an open K22). The number of occurrences of the pattern in a sample, X, is given by $X = \sum_{A \in \mathcal{A}} X_A$, where X_A is the random variable which is equal to 1 if all the arcs of pattern A are selected in the sample and 0 otherwise.

If we note l the number of arcs of the pattern (4 for a K22 and 3 for an open K22), we have that $\mathcal{P}[X_A = 1] = p^l$. By linearity of the expectation, we get $\mathbb{E}[X] = p^l|\mathcal{A}|$. Thus, $Y = p^{-l}X$ is an unbiased estimator of $|\mathcal{A}|$.

Variance. Note that the random variables X_A are not independent, i.e., two K22s can share a common link. Otherwise, the variance would simply be $\mathbb{V}(X) = \sum_{A \in \mathcal{A}} \mathbb{V}[X_A] = |\mathcal{A}|p^l(1 - p^l) \leq |\mathcal{A}|p^l$. However, we can argue that (and we will verify that), in practice, most of the K22s and open K22s do not share any link. It can be used in the analysis as follows.

$$\mathbb{V}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \mathbb{E}[(\sum_{A \in \mathcal{A}} X_A)^2] - \mathbb{E}[X]^2 \tag{3.1}$$

$$= \sum_{(A,B) \in \mathcal{A}} \mathbb{E}[X_A X_B] - \mathbb{E}[X]^2 \tag{3.2}$$

We now distinguish the couples of dependent patterns, which we note $\Delta = \{(A, B) \mid$

$A \cap B \neq \emptyset\}$, from the ones of independent ones, $\bar{\Delta} = \{(A, B) \mid A \cap B = \emptyset\}$.

$$\mathbb{V}[X] = \sum_{(A,B)\in\bar{\Delta}} \mathbb{E}[X_A X_B] + \sum_{(A,B)\in\Delta} \mathbb{E}[X_A X_B] - \mathbb{E}[X]^2 \tag{3.3}$$

When A and B are independent, we have

$$\mathbb{E}[X_A X_B] = \mathbb{E}[X_A]\mathbb{E}[X_B] = p^{2l}.$$

As $\mathbb{E}[X]^2 = p^{2l}|\mathcal{A}|^2$, we get

$$\mathbb{V}[X] = \sum_{(A,B)\in\bar{\Delta}} \mathbb{E}[X_A]\mathbb{E}[X_B] + \sum_{(A,B)\in\Delta} \mathbb{E}[X_A X_B] - \mathbb{E}[X]^2 \tag{3.4}$$

$$= \sum_{(A,B)\in\Delta} (\mathbb{E}[X_A X_B] - p^{2l}) \tag{3.5}$$

Let us now distinguish different cases. We note Δ_i the set of couples of patterns sharing $1 \leq i \leq l$ arcs. For a couple $(A, B) \in \Delta_i$, we have that $\mathbb{E}[X_A X_B] = p^{2l-i}$, giving that

$$\mathbb{V}[X] \leq \sum_{i=1}^{l} \sum_{(A,B)\in\Delta_i} (p^{2l-i} - p^{2l}). \tag{3.6}$$

Since $p < 1$, we get

$$\mathbb{V}[X] \leq \sum_{i=1}^{l} p^{2l-i} \left|\Delta_i\right|. \tag{3.7}$$

Note that, when all patterns are independent, $|\Delta| = |\Delta_l| = |\mathcal{A}|$ (couples $(A, A) \in \mathcal{A}$), giving back the variance of the independent case, $p^l|\mathcal{A}|$. Chebycheff's inequality tells us that:

$$Prob[|Y - \mu| \geq k\sigma] \leq \frac{1}{k^2}, \tag{3.8}$$

where μ is the expectation and σ is the standard deviation of X. In our case, if we want an accuracy of ε with a probability q, we should have $\frac{1}{k^2} \leq 1 - q$ and $k\sigma \leq \varepsilon p^l|\mathcal{A}|$, which can be rewritten as:

$$\frac{k^2}{\epsilon^2} \sum_{i=1}^{l} p^{2l-i} \frac{|\Delta_i|}{|\mathcal{A}|^2} \leq p^{2l}. \tag{3.9}$$

Lastly, to estimate the `icc`, we use as an estimator

$$Z = \frac{4Y}{Y_0}, \tag{3.10}$$

with Y and Y_o the estimators of the number of K22s and open K22s, respectively. As $\lim_{n\to\infty} Y = \#K22s$ and $\lim_{n\to\infty} Y = \#openK22s$, we have that $\lim_{n\to\infty} Z =$ `icc`. For the precision, if Y and Y_o have an accuracy of ε and ε_o respectively, then with a probability $q = 0.99$, Z has at least an accuracy of $\frac{1+\varepsilon}{1-\varepsilon_o} \underset{\varepsilon\to 0}{\sim} 1 + \varepsilon + \varepsilon_o$ with a probability $q^2 \approx 0.98$.

Numerical application. We now consider the K22s of the TS. Note that we know that $\frac{|\Delta_4|}{|\mathcal{A}|^2} = 1/\#K22s = 3.8 \times 10^{17}$. We also can notice that $|\Delta_3 = \Delta_4|$. In the TS,

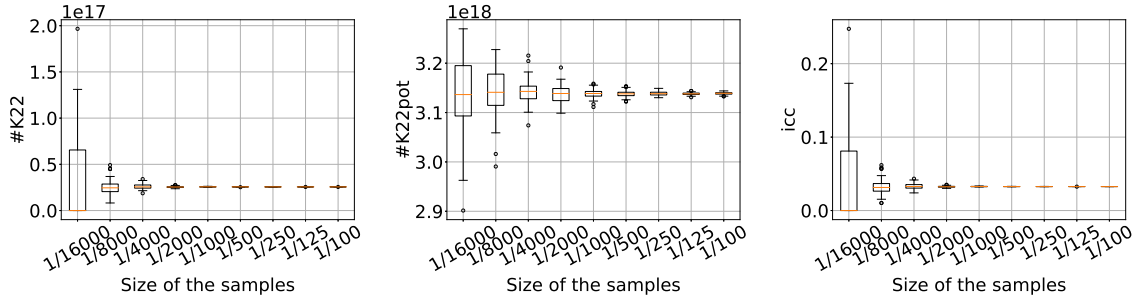


Figure 3.3: Estimation of the K22s (Left), open K22s (Middle) and interest clustering coefficient (Right) for different sample sizes.

an edge is shared by $\frac{\#K22s}{m}$ K22s on average, with m the number of links of the TS. Thus, the average number of K22s sharing at least an edge with a K22 is between $\frac{\#K22s}{m}$ and $4 \cdot \frac{\#K22s}{m}$. It gives $\frac{1}{m}|\mathcal{A}|^2 \leq \Delta_1 + \Delta_2 + \Delta_3 + \Delta_4 \leq \frac{4}{m}|\mathcal{A}|^2$. The number of overlapping K22s with i arcs is a non-increasing function of i. To make a numerical evaluation, we suppose that most overlapping K22s share one edge and not 2 edges in the TS. We set that $|\Delta_1| = \frac{1}{m} = 4.3 \times 10^{-8}|\mathcal{A}|^2$, and $|\Delta_2| = 10^{-16}|\mathcal{A}|^2$. Now, if we want a precision of $\varepsilon = 0.1$ with a probability 0.99 (that is $k = 10$), we need to take a sampling probability p such that

$$p^8 \geq \frac{10^2}{10^{-4}}(p^7 4.3 \times 10^{-8} + p^6 \times 10^{-16} + p^5 3.8 \times 10^{-17} + p^4 3.8 \times 10^{-17}). \quad (3.11)$$

That is $p \geq 2.5 \times 10^{-4}$. Thus, under these hypotheses, a sample with sampling probability 1/2500 and larger, e.g., our 1/2000 sample, allows to estimate the number of K22s with a precision of 10%. The number of open K22s is larger and thus, the precision is better. It gives a precision of at least $\frac{1+1/100}{1-1/100} = 0.20$ for the estimation of `icc`. In practice, the Chebysheff inequality and our hypothesis are pessimistic as shown below.

Results. We present in Figure 3.3 the results of the algorithm for different sample sizes, corresponding to sampling probabilities from $p = 1/100$ to p=1/16,000. For each sample size, we generated 30 samples. The distribution over the samples of the interest clustering coefficient, K22s and open K22s are provided by a boxplot for each value of p. Note that a K22 of the TS appears in a sample with a probability of only p^4, and of p^3 for an open K22. The clustering coefficient of a sample is thus an estimate of $p \cdot$ `icc`.

We observe that the clustering coefficient is well estimated using any sample for a sampling probability of 1/1000 or larger. Indeed, for this range of probabilities, the distribution over all samples is very concentrated and around the exact value of the `icc`. Note that, for $p = 1/1000$, a K22 is present in the sample with a probability of only 10^{-12}. The expectation of the number of nodes with an edge is only 23 million nodes (over 500 million) and the number of edges also around 23 million. Thus, a small sample (5% of the nodes and 0.1% of edges) allows to do an efficient estimation of the `icc`.

For smaller values of p, the variance increases. The median estimates well the `icc` for a range of p between 1/8000 and 1/1000, but samples of these sizes may have error of 100% of the value. Lastly, for $p = 1/16000$, only the number of open K22s (and not the K22s or the `icc`) is approximated by the median.

In conclusion, a sample with sampling probability 1/1000 is enough to efficiently

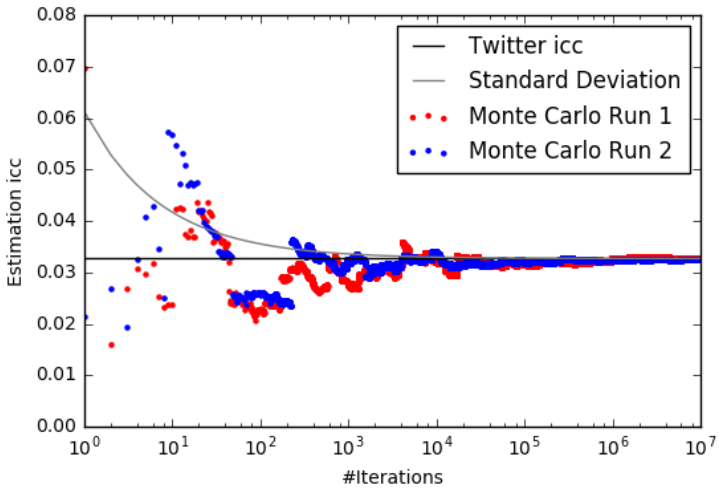


Figure 3.4: Estimation of the clustering coefficient with Monte Carlo Method.

estimate the interest clustering coefficient, with a computation time of around 1 minute (instead of days for the whole TS) on a machine of the cluster.

Monte Carlo Method.

After a short reminder of the precision of the Monte Carlo Method, we first quickly discuss the case of triangles to show the particularity of estimating the interest clustering coefficient. The difficulty here is that the probability to observe a (closed or open) K22 or a triangle is very small. In the case of triangles, this difficulty can be easily circumvented by knowing the node degrees. This allows to select an open triangle uniformly at random. In the case of K22s, this information is not sufficient to select an open K22 uniformly at random. In fact, achieving this goal is very costly, but we present a method in which, by picking only forks (as we do for triangles), we can compute the interest clustering coefficient.

Preliminary: Precision of Monte Carlo Method. *Precision of the estimation and number of iterations.* Each trial is a Bernoulli variable with probability p. We use as an estimate Y, the mean of the random sample. Its expectation is p and its standard deviation is $\frac{\sqrt{p(1-p)}}{\sqrt{n}}$. Due to the central limit theorem, we get that, when n is large,

$$Prob\left[|Y-p| \leq Z_{\alpha/2}\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right] = \alpha, \tag{3.12}$$

with $Z_{\alpha/2}$ the value giving the α confidence interval a standard normal distribution. To get with probability α an accuracy of ε of the empirical mean p (which is not known), we should have $Z_{\alpha/2}\frac{\sqrt{p(1-p)}}{\sqrt{n}} \leq \varepsilon p$. That is $n \geq \frac{Z^2_{\alpha/2}(1-p)}{p\varepsilon^2}$. If we take $n \geq \frac{Z^2_{\alpha/2}}{p\varepsilon^2}$, we have the wanted precision (and we are not doing many more iterations when p is small). For example, to get an accuracy of 99% ($\varepsilon = 0.01$), with probability $\alpha = 0.99$, we should have a number of iterations such that $n \geq \frac{75{,}625}{p}$.

Approximating the number of undirected triangles. A first direct method

would be to select three vertices uniformly at random and check if they form a triangle and open triangles. The problem with this method is that the probability to form a triangle in Twitter is the number of triangles divided by the number of triplet of nodes, i.e., $\frac{6.23\times10^{11}}{(5\times10^8)^3} = 5\times10^{-15}$. Thus the number of needed iterations would be astronomic, 5.5×10^{19} for an accuracy of 1%, with probability $\alpha = 0.99$. We thus have to use methods selecting open triangles directly.

To estimate the undirected clustering coefficient, we need to select open (undirected) triangles uniformly at random. We then test if the selected triangle is closed or not (which is the case with probability ucc). The number of open triangles rooted at vertex v is equal to $\frac{d(v)d(v)-1}{2}$. We can thus perform the sampling by picking a vertex v with probability $\binom{d(v)}{2}/\sum_{v\in V}\binom{d(v)}{2}$ and then select two random edges adjacent to v.

Directed triangles. The method is the same in the case of directed triangles. We select an open triangle uniformly at random. The number of open triangles rooted at a vertex v is $d^-(u)d^+(u)$. We thus select a node u with probability $d^-(u)d^+(u)/\sum_{v\in V} d^-(v)d^+(v)$. We then select uniformly at random an incoming arc and an outgoing arc. Lastly, we check if the triangle is closed (which is the case with a probability equal to tcc and to ccc respectively for transitive and cyclic triangles).

Precision of the estimation and number of iterations. Each trial is a Bernoulli variable with a probability $p = tcc = 0.019$. To get an accuracy of 1%, with probability 0.99, we should thus do $n = 4\times10^6$ iterations.

Interest clustering coefficient. For triangles, we were able to select uniformly at random open triangles using the node degrees. In the case of K22s, node degrees is not sufficient to select an open K22 uniformly at random. To do so, it would be necessary to compute the number of open K22s with u as a root. This preprocessing is very costly: for each node, we should consider its in-neighbors, sum their out-degrees, and compute the number of internal edges. It would be almost as costly as doing an exact count of the number of K22s.

Another method is to select a vertex v as a root according to the square of its in-degree (as in the case of triangles), but without knowing its number of open K22s (first step). We then select two arcs u_1v and u_2v uniformly at random (second step). We then compute the number of K22s and open K22s with the selected fork (u_1v, u_2v) (third step).

For the first step, the algorithm needs a list of the node in-degrees of the TS, which would have been computed in a preliminary step. For the second one, it then uses the in-adjacency of v. For the third step, the out-adjacency of u_1 and u_2 are necessary for the computations.

We then use the estimators introduced below. We first define

$$X = \#K22s(u_1v, u_2v) \quad \text{and} \quad X_o = \#openK22s(u_1v, u_2v).$$

We have

$$\mathbb{E}[X] = \sum_{forks} \#K22s(fork)\mathbb{P}(fork). \tag{3.13}$$

As each fork is chosen uniformly at random and as a K22 has two forks, we get

$$\mathbb{E}[X] = \sum_{forks} \#K22s(fork)\frac{1}{\#forks} = \frac{2\#K22s}{\#forks}. \tag{3.14}$$

	$\#closed$	$\#open$	cc
`icc`	$25,605,832,012,451,571$	$3,138,466,676,914,054,233$	0.032634831
	2.6×10^{16}	3.1×10^{18}	3.3%
tcc	$2,469,018,039,988$	$129,023,573,841,024$	0.019136178
	2.5×10^{12}	1.3×10^{14}	1.9%
ccc	$723,131,368,202$	$129,023,573,841,024$	0.016813936
	7.2×10^{11}	1.3×10^{14}	1.7%
ucc	$623,873,346,660$	$1,631,948,600,661,523$	0.001146862
	6.2×10^{11}	1.6×10^{15}	0.11%
mcc	$317,649,850,664$	$8,924,125,201,234$	0.106783526
	3.2×10^{11}	8.9×10^{12}	10.7%

Table 3.1: Clustering coefficients (exact and approximated count) in the TS.

Similarly,

$$\mathbb{E}[X_o] = \frac{\#openK22s}{\#forks}. \tag{3.15}$$

We may thus define two efficient unbiased estimates for #K22s and #openK22s:

$$Y = \frac{\#forks}{2n} \sum_{i=1}^{n} X_i. \quad \text{and} \quad Y_o = \frac{\#forks}{n} \sum_{i=1}^{n} X_{oi}. \tag{3.16}$$

We have $\mathbb{E}[Y] = \#K22s$ and $\mathbb{E}[Y_o] = \#openK22s$. The number of forks with a vertex v as a root is given by $\binom{d^-(v)}{2}$. The total number of forks in the TS is thus $\sum_{v \in V} \binom{d^-(v)}{2}$. Lastly, as we are interested in the interest clustering coefficient, we define

$$Z = \frac{4Y}{Y_0}. \tag{3.17}$$

As $\lim_{n\to\infty} Y = \#K22s$ and $\lim_{n\to\infty} Y = \#openK22s$, we have that $\lim_{n\to\infty} Z =$ `icc`.

Experiments. We carried out two runs with 10 million iterations. It took about 2min30 for one run (60,000 iterations per second). The value of the estimator of the `icc` for the two runs is plotted as a function of the number of iterations in Figure 3.4. We first see that the estimator converges as expected to the value of the `icc` of TS represented by a straight horizontal line (and which was computed exactly in the previous section). We also plotted the estimated standard deviation as a function of the number of iterations. To obtain it, we did one billion iterations. We then estimated the standard deviation σ, and plotted $\frac{\sigma}{\sqrt{n}}$. We see that large jumps or discontinuity happen, but only at the beginning. They correspond to the draw of a fork with a lot of K22s and open K22s corresponding to a user who does not have the same `icc` as the global network. Then, the convergence is quick. After 300 iterations, the standard deviation is below 10% and after 1000 iterations, we do not experience a value of the runs less precise than 10%.

3.4 Results: Clustering coefficients in Twitter

To compute the number of K22s and open K22s, directed triangles, and undirected triangles in the Twitter Snapshot, we used a cluster with a rack of 16 Dell C6420 dual-Xeon 2.20GHz (20 cores), with 192 GB RAM, all sharing an NFS Linux partition over Infiniband. It took 51 hours to compute the exact numbers of K22s and open K22s, corresponding to 265h of cumulative computation times on the cluster. We reported the results in Table 3.1.

Number of K22s and triangles. We see that the numbers of K22s and open K22s are huge, 2.6×10^{16} and 3.1×10^{18}, respectively. It has to be compared with the number of triangles which are several orders of magnitude smaller: e.g., 2.5×10^{12} and 1.3×10^{14} for transitive triangles.

Clustering coefficient in the mutual graph. The mutual graph captures the friendship relationships in the social network. The mutual clustering coefficient thus is high ($mcc = 10.7\%$), as cliques of friends are frequent in Twitter.

Clustering coefficients in the whole graph. We observe that $\texttt{icc} = 3.3\% > tcc = 1.9\% > ccc = 1.7\% > ucc = 0.11\%$. Directed metrics better capture the interest relationships in the TS as ucc is very low. The highest parameter is the `icc`. It confirms the hypothesis of this Chapter that common interests between two users are better captured by the notion of K22 than by a direct link between these users. As expected, the second parameter is the one using transitive triangles. Indeed, they capture a natural way for a user of finding a new interesting user, that is, considering the followings of a following, especially after having seen retweets. A bit surprisingly, the ccc is not very low. In fact, a large fraction of the cyclic triangles are explained by corresponding triangles in the mutual graph (triangles of bi-directional links).

A way to artificially take off the social influence in order to focus exclusively on the directed interest part of the graph is to remove the (open and closed) triangles and K22s contained in the mutual graph from the total count. Indeed, each undirected triangle of the mutual graph induces two cyclic triangles and four transitive triangles, and each undirected open triangle induces two open triangles. In the same way, each undirected K22 induces two K22s and each undirected open K22 induces two open K22s. The obtained results are shown in Table 3.2. If we take off those mutual triangles, both the tcc and the ccc values drop to 0.51% and 0.24%, respectively, while the `icc` stays about the same at 3.1%. This tends to confirm the hypothesis that the directed triangle clusterings somehow measure the friendship part of the TS more than the interest part.

We can even go one step further by computing the number of triangles in the graph in which all bidirectional edges have been removed. In that case, the *ccc* drastically drops to 0 (we found no cyclic triangles without at least a bidirectional arc in the dataset!) while tcc and interest clustering coefficient stay almost the same, 3.6 and 4.2 respectively. This confirms that cyclic triangles are artificially created by friendship relations and that the *ccc* gives no information about the directed part of the graph.

Distribution of the `icc` and local clustering. We also provide the distribution of the values of the interest clustering coefficient over all users (having open K22s) in Figure 3.5. We see that the `icc` greatly varies between 0 and 1. A large number of nodes have a low value of `icc`, e.g., 2.23×10^7 users (10.2% of the users with

	icc	*tcc*	*ccc*	*ucc*
Twitter	3.1%	0.51%	0.24%	0.057%

Table 3.2: Clustering coefficients without the mutual structures.

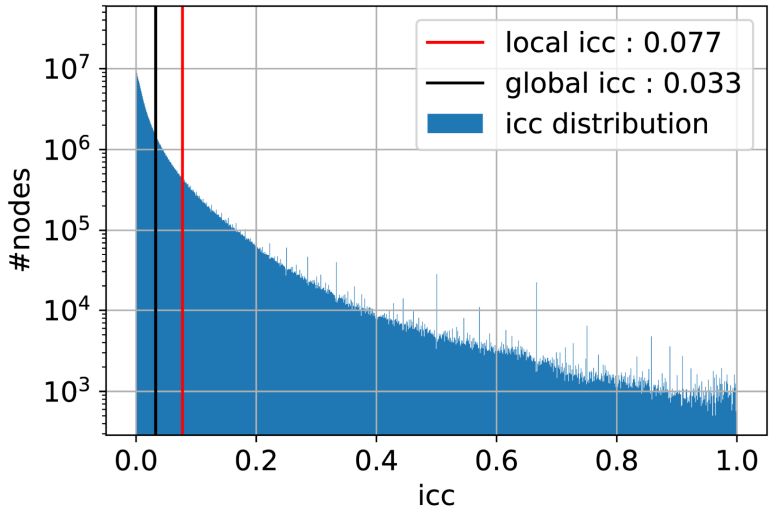


Figure 3.5: Histogram of the distribution of the interest clustering coefficient over all users of the Twitter Snapshot. The vertical bars indicate the value of the glocal `icc` (3.3%) and the average value (7.7%) or local `icc`.

open K22s) have a value of 0, meaning they are part of open K22s but not of K22s. At the opposite end, 2.4×10^4 users (0.011% of the users with open K22s) have a value of 1, meaning that all their open K22s belong to a K22. The *average value* is equal to 7.7%. This value could be used as a definition of a *local* $\texttt{icc}^{\top}$. Indeed, as discussed above, the number of K22s and open K22s per user have been computed while considering a user as a top vertex. A second local coefficient, $\texttt{icc}_{\perp}$, can be defined for bottom vertices.
Similarly to what was found in Facebook, the local coefficient has a larger value than the global one. This may be due to the fact that a large number of nodes with few K22s and open K22s (usually nodes with small degrees) only are in a single small strongly connected community, and thus have a higher than average `icc`. On the contrary, a small number of nodes with larger degrees and larger number of K22s and open K22s may be in different communities, leading to smaller than average `icc`.

3.5 Results: Other Directed Datasets

We computed the different metrics on four other directed networks: two social networks, a web network and a citation network. The data information are gathered in Table 3.3, while the clustering coefficients are reported in Table 3.4. We also computed the values of the clustering coefficients without the mutual structures (not provided here); interestingly, those values are close to the ones on the total graphs.

	Is a Social Network	N	$\|E\|$	$\frac{\|E\|_m}{\|E\|}$
Instagram	Yes	4.5×10^4	6.7×10^5	11%
Flickr	Yes	2.3×10^6	3.3×10^7	62%
Web (.edu)	No	6.9×10^5	7.6×10^6	25%
Citations	No	3.8×10^6	1.7×10^7	0%

Table 3.3: Datasets information. N is the number of nodes, $|E|$ the number of edges, and $\frac{|E|_m}{|E|}$ the fraction of edges implied in a bidirectional link.

	icc	tcc	ccc	mcc	ucc
Instagram	12.0%	15.4%	3.7%	22.6%	4.1%
Flickr	12.4%	12.2%	9.3%	13.9%	10.8%
Web (.edu)	46.3%	59.6%	18.8%	78.5%	0.69%
Citations	22.3%	9.1%	0%	*(none)*	6.7%

Table 3.4: Clustering coefficients of the directed datasets.

We observe that the structure of each dataset is revealed by (the mix of) values of the different clustering coefficients, as discussed below.

Instagram: Instagram is a photo and video-sharing social network. This dataset was collected by Ferrara et al. [12] in 2014. The network is close to the Twitter one. Nodes corresponds to the accounts, and there is a link $u \to v$ if the account u follows the account v. The results are quite similar to what we found for Twitter: the *icc* and *tcc* are high and of the same order; the *ccc* is also high because of the bidirectionnal edges (it drastically drops to 0.06% when removing those links). The *mcc* is the highest value, while the *ucc* is lower than the others. This confirms that social networks share some common characteristics.

Flickr: Flickr is an image and video hosting service, which allows you to follow other people on the plateform to see more easily their content. The dataset was collected in 2008 by Mislove et al. [25]. This is once again a graph of followers of a directed social network. The values are similar to the previous one but for the *ucc*, which is higher. We can notice that Flickr looks more like a social media than Twitter and Instagram, since there is 62% of links implied in bidirectional. This explains why the undirected clustering coefficient is not so different from the mutual one .

Berkley-Stanford.edu web pages: The dataset was collected in 2002 by Leskovec et al.[23]. The nodes represent the pages from berkely.edu and stanford.edu domains and directed edges represent hyperlinks between them. The *tcc*, *icc*, and *mcc* are really high. For the *tcc*, this is due to the very hierarchical structure of the institution web pages. As an example, a researcher will be linking towards his group, laboratory, and university in its website, while the group website is linking to its laboratory and university... This strong structure translates into a high value of the *tcc*. As for the *icc*, research and educational domains form naturally strong communities creating large number of common neighbors for individuals of the same domain, and thus

a high *icc*. Groups/teams/departments also constitutes strong social communities, leading to a high *mcc*.

Citations: Collected by Leskovec et al.[22], it includes all citations made by patents granted between 1975 and 1999. This is a good example of information network, giving a high value of `icc` of 22.6%, while the tcc value is 9.1%. Indeed, research fields and industry domains are strong communities leading to a high `icc`. Moreover, it is also not rare to cite a patent and its citations (the patent acting as a survey), explaining the tcc value. Note that there are no cyclic triangles nor bidirectional links, because of the temporal structure of citations - a paper will only cite older papers.

Takeaways: The following takeaways summarize the variety of informations given by the different clustering coefficients:

- A high value of `icc` indicates the presence of clusters of interests such as research communities or interest fields.
- A high value of tcc is the sign of an important *local* phenomena of friends' or acquaintances' recommendations and/or of a high hierarchical structure in the dataset.
- The ccc has no real social meaning. If its value can be high in a directed graph, this is only due to the presence of bidirectional arcs and triangles. The closure of a cyclic triangles is very rare in directed networks with no bidirectionnal edges, confirming the general intuition.
- Directed networks have a high *mcc*. Indeed, their bidirectional parts (mutual graph) have strong social communities, leading to a high clustering coefficient.
- The ucc is usually significantly lower, showing that the directed part of the network is better understood using directed clustering coefficients.
- Directed social networks have similar mixes of values of their undirected and directed clustering coefficients, however, with some notable differences, due to their diverse usages and information.

3.6 Model with addition of K22s

To model complex networks, a model with a high number of triangles was introduced in [38]. In this section, we introduce a new random graph model in which the *number of K22s is higher* than classical directed random graphs. The model is based on the model from Bollobás et al. [5] to which we add what we call a K22 event. A K22 event closes an open K22. The principle is that if a user has a common interest with another user, and if this user has another interest, it has an increased chance to be interested and to follow it. We then show that the in-degree and out-degree distributions of the introduced model follow a power law (as many real networks). Lastly, we exhibit the increase of the interest clustering coefficient of the generated graphs with the probability of a K22 event.

3.6.1 Presentation of the model

We recall here the events defining the classic preferential attachment model of [5] and define the K22 event. We start with an initial graph $G_0 = (V_0, E_0)$. Then, at each time step t:

- With a probability (1-p) (**Bollobás et al. event**):
  - With a probability α, we add a node u and a link leaving this node and reaching an existing node v chosen with a probability proportional to $d_{in}(v) + \delta_{in}$;
  - With a probability β, we add a node v and a link reaching this node and leaving an existing node u chosen with a probability proportional to $d_{out}(u) + \delta_{out}$;
  - With a probability $1-\alpha-\beta$, we add an edge between two existing nodes, chosen with probability proportional to $d_{out}(u)+\delta_{out}$ for the leaving node u and $d_{in}(v) + \delta_{in}$ for the reached node v.
- With a probability p (**K22 event**):
  1) We choose a random node (called u_1) with a probability proportional to its out-degree $d_{out}(u_1)$;
  2) We pick uniformly at random an out-neighbor of the node u_1 (called v_1);
  3) We pick uniformly at random an in-neighbor of the node v_1 (called u_2);
  4) We pick uniformly at random pick an out-neighbor of the node u_2 (called v_2);
  5) We add a link from u_1 to v_2.

The idea of the K22 event is to close an open K22; since u_2 follows v_1 and v_2 at the same time, v_1 and v_2 have a higher probability to be similar, and a person u_1 following v_1 has a higher chance to be interested in v_2.
Note that it is possible to introduce multiedges with the K22 events. Indeed, to make the problem tractable, we allow $u_1 = u_2$ in Step 3, or $v_2 = v_1$ in Step 4. In the empirical study, we construct the random graphs with the multiedges and we get rid of them at the end of the constructions. We empirically verify that the multiedges do not impact the results in the end of the section. Indeed, most of them appear for low degree nodes and, thus, they do not affect the tail of the degree distributions.

3.6.2 In-degree and out-degree distributions

We show in what follows that the in- and out-degree distributions of the introduced model follow power-laws, as most real networks. More precisely:

Theorem 1. *The probability $P(i)$ (resp. $P(o)$) for a node to have in-degree i (resp. out-degree o) in the new model is:*

$$P(i) \underset{i>>1}{\sim} i^{-(1+\frac{1}{A})} \quad \text{and} \quad P(o) \underset{o>>1}{\sim} o^{-(1+\frac{1}{B})},$$

where $A = p + \frac{(1-p)(1-\beta)}{1+(1-p)(\alpha+\beta)\delta_{in}}$ *and* $B = p + \frac{(1-p)(1-\alpha)}{1+(1-p)(\alpha+\beta)\delta_{out}}$.

Proof. We first focus on the in-degree distribution. This result is derived from the equation giving the evolution of the number of nodes of in-degree i as a function of time, sometimes called Master Equation.
Let $G(t) = (V(t), E(t))$ be the graph obtained at time t, and $N(t) = |V(t)|$. The number of edges at time t is $|E(t)| = t + |E_0| \approx t$, while the number of nodes is $N(t) = (1-p)(\alpha+\beta)(t+|V_0|) \approx (1-p)(\alpha+\beta)t$ when t is high enough. Hence, the mean in-degree (and out-degree) of the network is $m = \frac{1}{(1-p)(\alpha+\beta)}$.

Let us compute the in-degree distribution. Calling $N(i,t)$ the number of nodes of in-degree i at time t, we can write the Master Equation:

$$N(i,t+1) - N(i,t) = (1-p)\alpha\delta_{0,i} + (1-p)\beta\delta_{1,i} \tag{3.18}$$

$$+ (1-p)(1-\beta)\frac{i-1+\delta_{in}}{\sum_{i=0}^{+\infty} N(i,t)(i+\delta_{in})}N(i-1,t) \tag{3.19}$$

$$- (1-p)(1-\beta)\frac{i+\delta_{in}}{\sum_{i=0}^{+\infty} N(i,t)(i+\delta_{in})}N(i,t) \tag{3.20}$$

$$+ p\frac{i-1}{\sum_{i=0}^{+\infty} N(i,t)i}N(i-1,t) - p\frac{i}{\sum_{i=0}^{+\infty} N(i,t)i}N(i,t) \tag{3.21}$$

where $\delta_{i,j}$ is the Kronecker delta.
The Master Equation formulates the variation of the number of nodes with degree i between time i and time $i+1$. The two first terms on the right hand side correspond to the addition of a new node, with degree 0 or 1 (depending on if we are in the first or second case of the Bollobás et al. event). The third and fourth terms are the probabilities that, during the Bollobás et al. event, an edge is connected to a node of degree $(i-1)$ or i. This would lead to the arrival of a new node of degree i, or the loss of one of them. Those events occur with probability $(1-p)(\alpha+(1-\alpha-\beta))$. Finally, the last two terms correspond to the probability that an edge connects a node of degree $(i-1)$ or i during the K22 event.
We now show that the probability to connect to a node (v_2) of a given degree after following an open K22 is proportional to the degree of this node. Indeed, the probability to connect to a node (v_2) of a given degree after following an open K22 is

$$P(x = v_2) = \sum_{y \in N^+(v_2)} P(y = u_2) \times \frac{1}{d_{out}(y)}, \tag{3.22}$$

where $N^+(v_2)$ is the set of in-neighbors of v_2, and u_2 is defined in the model. Using the same reasoning, we have

$$P(x = u_2) = \sum_{y \in N^-(u_2)} P(y = v_1) \times \frac{1}{d_{in}(y)} \tag{3.23}$$

and

$$P(x = v_1) = \sum_{y \in N^+(v_1)} P(y = u_1) \times \frac{1}{d_{out}(y)}. \tag{3.24}$$

Since $P(y = u_1) = \frac{d_{out}(y)}{t}$, we deduce that

$$P(x = v_2) = \frac{d_{in}(x)}{t}, \tag{3.25}$$

which gives us the expected result.
Using this property and knowing that

$$\sum_{i=0}^{+\infty} i \cdot N(i,t) = |E(t)| = t \tag{3.26}$$

and

$$\sum_{i=0}^{+\infty} N(i,t)\delta_{in} = \delta_{in} N(t) = (1-p)(\alpha+\beta)\delta_{in}, \tag{3.27}$$

we can rewrite the equation as:

$$N(i,t+1) = \alpha\delta_{0,i} + \beta\delta_{1,i} \tag{3.28}$$

$$+ \Big(p\frac{i-1}{1} + (1-p)(1-\beta)\frac{i-1+\delta_{in}}{1+(1-p)(\alpha+\beta)\delta_{in}}\Big)\frac{N(i-1,t)}{t} \tag{3.29}$$

$$- \Big(1 + \big(p\frac{i}{1} + (1-p)(1-\beta)\frac{i+\delta_{in}}{1+(1-p)(\alpha+\beta)\delta_{in}}\big)\frac{1}{t}\Big)N(i,t). \tag{3.30}$$

Let us call

$$Z \equiv 1 + (1-p)(\alpha+\beta)\delta_{in}. \tag{3.31}$$

We need the following lemma from [10]:

Lemma 3 ([10])**.** *If we have an equation of the form :*

$$N(i,t+1) = \Big(1 - \frac{b(t)}{t}\Big)N(i,t) + g(t) \tag{3.32}$$

where $b(t) \to b$ and $g(t) \to g$ as $t \to +\infty$, then

$$\frac{N(i,t)}{t} \to \frac{g}{b+1}. \tag{3.33}$$

Using Lemma 3 and calling $P(i) = \lim\limits_{t\to+\infty} \frac{N(i,t)}{t}$, we have:

$$P(i) \;\; = \frac{\Big(\frac{(1-p)(1-\beta)}{Z}+p\Big)(i-1)+\frac{\delta_{in}}{Z}}{1+\Big(\frac{(1-p)(1-\beta)}{Z}+p\Big)i+\frac{\delta_{in}}{Z}} P(i-1). \tag{3.34}$$

Let us call

$$A \equiv \frac{(1-p)(1-\beta)}{Z} + p. \tag{3.35}$$

We thus have:

$$P(i) \quad = \frac{i-1+\frac{\delta_{in}}{ZA}}{i+\frac{\delta_{in}}{ZA}+\frac{1}{A}} P(i-1) \tag{3.36}$$

$$= P(1)\prod_{k=2}^{i} \frac{k-1+\frac{\delta_{in}}{ZA}}{k+\frac{\delta_{in}}{ZA}+\frac{1}{A}} \tag{3.37}$$

$$= \frac{\Gamma(i+\frac{\delta_{in}}{ZA})\Gamma(\frac{1}{A}+\frac{\delta_{in}}{ZA}+2)}{\Gamma(i+\frac{\delta_{in}}{ZA}+\frac{1}{A}+1)\Gamma(\frac{\delta_{in}}{ZA}+1)}. \tag{3.38}$$

Leading to

$$P(i) \underset{i>>1}{\sim} i^{-(1+\frac{1}{A})}. \tag{3.39}$$

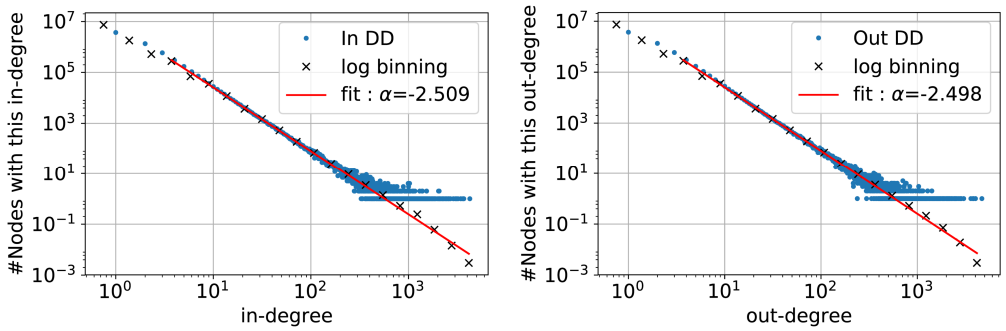


Figure 3.6: In- (Left) and out- (Right) degree distributions of a network built with the new model. The obtained distribution is given by the blue points; the black crosses represent the logarithmic binning of the distribution (a mean of a given amount of points on a logarithmic scale). The red straight line is the fit of the logarithmic binning; it has slope of -2.509 (resp. -2.498) for the in- (resp. out-) degree distribution (expected slopes from analysis are -2.5).

The *out-degree distribution* calculation follows the same method. The master equation is the same, except that δ_{in} and β are replaced by δ_{out} and α. The slope of the out-degree distribution is thus:

$$P_{out}(o) \underset{o>>1}{\sim} o^{-(1+\frac{1}{B})}, \text{with } B = \frac{(1-p)(1-\alpha)}{1+(1-p)(\alpha+\beta)\delta_{out}} + p. \tag{3.40}$$

Concentration. We have studied here the mean of the distributions. We now use the Azuma's inequalities to show the concentration around the mean. We have the following result [10]: Let X_t be a martingale with $|X_s - X_{s-1}| \leq c$ for $1 \leq s \leq t$. Then:

$$P(|X_t - X_0| > x) \leq \exp(-x^2/2c^2t). \tag{3.41}$$

Let $Z(i,t)$ be the number of vertices of degree i at time t and let F_s denote the σ-field generated by the choices up to time s. We apply the result to $X_s = E(Z(i,t)|F_s)$. We have that $|X_s - X_{s-1}| \leq 2$. Indeed, when we add an edge in the network, we affect only the degrees of its two end-vertices. Since $Z(i,0) = E(Z(i,t))$, using the result with $x = \sqrt{t\log(t)}$, we have

$$P(|Z(i,t) - E(Z(i,t))| > \sqrt{t\log(t)}) \leq t^{-\frac{1}{8}}. \tag{3.42}$$

And hence, $\frac{Z(i,t)}{t} \underset{t\to+\infty}{\to} P(i)$ in probability. □

The degree distributions of the model follow power-laws, with exponents between -2 and $-\infty$. We notice that, for $p = 0$, we recover the exponents of the Bollobás et al. model $-(1 + \frac{1+(\alpha+\beta)\delta_{in}}{1-\beta})$ and $-(1 + \frac{1+(\alpha+\beta)\delta_{out}}{1-\alpha})$ [5], while, when p goes to 1, the exponent goes to -2.

Note that, similarly to the Bollobás et al. model, we cannot generate graphs with any wanted mean-degree and fixed slopes of the power-law. Some constraints exist in order to keep $\delta_{in} > 0$ and $\delta_{out} > 0$. For instance, with $\alpha = \beta = 0.4$ and slopes of -2.5 (the values of our experiments), p has to stay in the interval $[\frac{1}{6}, \frac{2}{3}]$.

Validation by simulations. We validate the analysis and the hypothesis by simulation. In Figure 3.6, we present the in- and out-degree distributions of a network

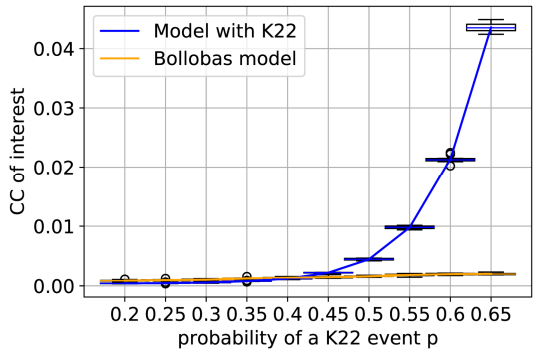


Figure 3.7: Interest clustering coefficient of our new model as a function of p, the probability of a K22 event. The value is compared with the one of the Bollobás et al. model [5].

built with our new model as an example. The parameters are fixed to $p = 0.5$, $\alpha = \beta = 0.4$, and $\delta_{in} = \delta_{out} = 2.0$. In this case, the expected slopes are -2.5. The fit is almost perfect: -2.509 and -2.498 for the in- and out-degree distributions.

3.6.3 Interest clustering coefficient of the new model

We show by simulation how the `icc` increases as p increases. We compare it with the one of the Bollobás et al. model. Note that, when p increases, the average degree of the model increases. Indeed, the mean degree is $m_{new} = \frac{1}{(1-p)(\alpha+\beta)}$. To compare networks with the same characteristics (mean degrees and exponents of the in-degree distribution), we adapt the parameters of the second model with the value of p.
Since, in the Bollobas et al. model, the mean degree is $m_{Bol} = \frac{1}{\alpha+\beta}$, we can compare the two models by: choosing the values of α, β, and p for our model. This imposes a value of m. We then choose α, β for the Bollobás et al. model, so that the two networks have the same mean degree. Finally, we choose δ_{in} so that the exponent of the in-degree distribution stays the same in both networks. In practice, we have fixed the exponent to -2.5 and imposed $\alpha_{new} = \beta_{new} = 0.4$. We compare the `icc` for both models for different values of p and report the results in Figure 3.7. We used graphs of size $N = 10^7$ nodes and averaged over 10 networks for each point. We see that the `icc` varies from 0.036% to 4.4% when p varies from 0.2 to 0.6.

3.7 Link Recommendation

We propose to use the K22s defined for our metric to carry out link recommendation, as we advocate that the interest clustering coefficient is a good measure of common user interests. For a neighbor, the principle is to recommend links closing open K22s. We define the *strength of a link* as the number of open k22s it would close if added to the graph. Links are then recommended by decreasing strengths. Typical recommendation systems propose the strongest link to a user (e.g., Facebook) or a top 10/top 20 list (e.g., Youtube).

We tested our method on the Twitter snapshot. We considered a population of 1000 users selected uniformly at random over the full population of Twitter's users.

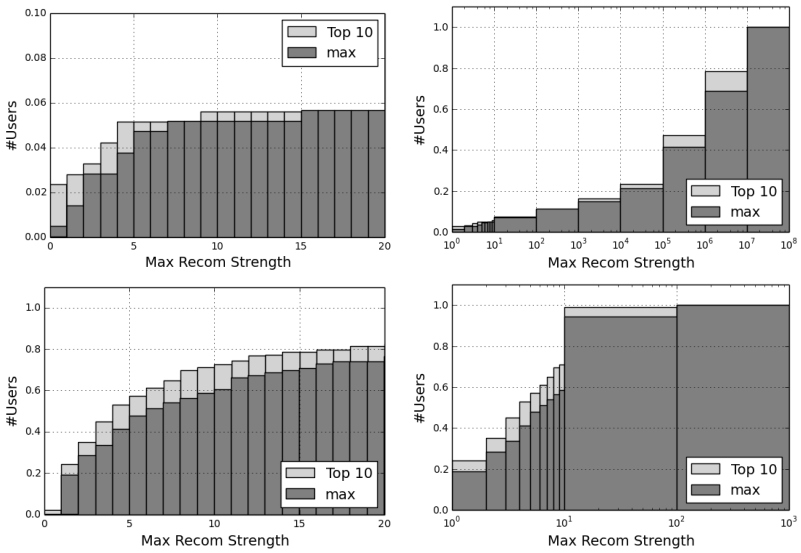


Figure 3.8: Cumulative distribution of the max and 10th recommendation strength over 1000 random Twitter's users for K22 recommendation (Top) and transitive triangle recommendation (Bottom). The left plots are a zoom on recommendations with weak strengths (≤ 20). The right plots present the complete cumulative distribution in log scale. Beware of the y-scale for the K22 zoom left plot.

Note that we excluded users following no one. Indeed, isolated users are not interesting users per se and for this study and they have no TT or K22 recommendations.

For each node, we computed its open K22s (for a node x, we follow all its out-neighbors, then for each out-neighbor, we follow its in-neighbors, then for each in-neighbor, we follow its out-neighbors. These last nodes (which were not already followed by x) are the recommended nodes. We then count how many times a node is recommended. This gives the link strength.

We compared the method with classic recommendations using triangles. For example, on Facebook, it is frequent to have a message such as "8 of your friends know Bob. Do you know Bob?" Connecting with Bob would close 8 open (undirected) triangles. As we are considering a directed graph and are focusing on interest links, we computed recommendations based on transitive triangles, as they have more social sense than cyclic triangles. For a user x, we recommend the out-neighbors of the out-neighbors of x.

Note that there are a lot more open K22s than open triangles in the graph, 3.1×10^{18} compared to 1.3×10^{14}. We argue in the following that it allows to make more recommendations and most importantly better recommendations.

We report in Figure 3.8 histograms of the cumulative distribution over the 1000 random users of the strengths of the recommendation with maximum strength and of the 10th recommendation. The top plots present K22 recommendations while the bottom ones the TT recommendations. The right plots show the complete cumulative distribution in log scale, while the left plots are a zoom on recommendations with weak strengths (≤ 20). Beware that the y-scale of the K22 zoom left plot which is between 0 and 0.1. Notice also the difference in x-scale for the right plots.

Top/Max recommendation. We remark that a small amount of users have TT recommendations and no K22 recommendation. This is due to the fact that for a

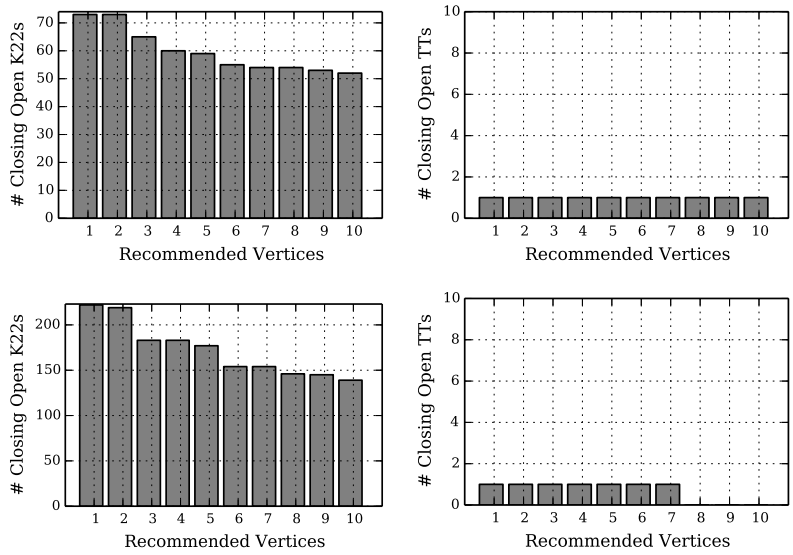


Figure 3.9: Strengths of the top 10 recommendations for 2 typical Twitter users using (Left) K22 recommendations (Right) TT recommendations.

user with few outgoing links, it is more probable that the followed users are also following at least one other user (providing a TT recommendation) than they are followed by other users (necessary to provide a K22 recommendation). We do not advocate to use only K22 recommendations, but to use it as a complementary tool. In particular, for users with no TT and K22, recommendations would only be made based on global social network statistics (trending topics for example).

However, when a K22 recommendation exists for a user, it has much more strength than the TT recommendations for her. Indeed, 21% of users have TT recommendations of strengths 0 or 1. This number is just 1.2% for K22 recommendations. A recommendation of strength 1 has very good chance to be of no interest, as it is based on the following of a single user over 500 million ones. Similarly, 28% of users only have TT recommendations of strengths 2 or lower (to be compared with 2.5% for K22 recommendations). This means that, for a very large portion of users, TT recommendations are based on very few links. On the contrary, more than 94% of users have a top K22 recommendation with strength more than 10. *We are thus able to carry out a meaningful recommendation for the vast majority of users using K22s.*

Top 10 recommendations. When considering a recommendation system proposing a top 10, we see that 25% of users have their 10th TT-recommendation of strength 1 or lower, and 35% of strength 2 or lower. There does not exist a significant top 10 list for more than one third of users. On the contrary, 94% of users have their 10th K22-recommendation with strength higher than 10. Top 10 recommendation systems can thus be implemented for most users using K22s. Moreover, the distribution of recommendation strengths is very flat when using TT (a large number of top recommendations have strength 1), see Figure 3.9. Thus, it is very hard to discriminate between recommended users and to do a meaningful ranking of recommendations. At the opposite end, the distribution usually is steep for K22. It is thus a lot easier to establish a ranking.

Typical users. We present in Figure 3.9 the strengths of the top 10 recommenda-

tions using K22 (Left) and TT (Right) for two typical users. For the first one (Top), it is implicated in around 200 triangles, representing each a potential recommendation. However, the strength of the recommendations is very low, just 1 for all of them. Recommendations for this user would be very bad for two reasons: first, they are based on the choice of only 1 user. Second, if the recommendation system had to propose a top 10, how would it discriminate between the 200 similar potential ones with similar strength. On the contrary, the K22 recommendations have much more strengths: 72 for the 1st and the 2d ones, and 52 for the 10th one. The K22 recommendations are thus much more well-grounded. For the second user (Bottom), we observe a similar phenomenon, but with fewer recommendations. It is not even possible to build a top 10 for her using TT as only 8 links can be proposed, and not with a high confidence (strength 1). Conversely, the top 10 K22 recommendations have strengths between 215 and 135.

3.8 Conclusion

In this Chapter, we introduce a new metric, the *interest clustering coefficient*, to capture the interest phenomena in a directed graph. Indeed, the classical undirected clustering coefficient apprehends the social phenomena that my friends tend to be connected. However, it is not adequate to take into account directed interest links. The interest clustering coefficient is based on the idea that, if two people are following a common neighbor, they have a higher chance to have other common neighbors, since they have at least one interest in common. We computed this new metric on a network known to be at the same time a social and information media, a snapshot of Twitter from 2012 with 505 million users and 23 billion links. The computation was made on the total graph, giving the exact value of the interest clustering coefficient, and using sampling methods. The value of the interest clustering coefficient of Twitter is around 3.3%, higher than (undirected and directed) clustering coefficients introduced in the literature and based on triangles, which we also computed on the snapshot. This consolidates the idea that Twitter is indeed used as a social and information media, and that the new metric introduced in this Chapter captures the interest phenomena. We then proposed a new model, building random directed networks with a high value of K22s, and a new method for link recommendation using K22s. As a future work, we would like to investigate further link recommendation based on the K22 structure defined for the interest clustering coefficient: in particular, it would be interesting to carry out a real-world user case study to investigate if users are more satisfied by such recommendations.

Bibliography

[1] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47, 2002.

[2] Noga Alon, Raphael Yuster, and Uri Zwick. Finding and counting given length cycles. *Algorithmica*, 17(3):209–223, 1997.

[3] Lars Backstrom, Paolo Boldi, Marco Rosa, Johan Ugander, and Sebastiano Vigna. Four degrees of separation. In *Proceedings of the 4th Annual ACM Web Science Conference*, pages 33–42. ACM, 2012.

[4] Alain Barrat and Martin Weigt. On the properties of small-world network models. *The European Physical Journal B-Condensed Matter and Complex Systems*, 13(3):547–560, 2000.

[5] Béla Bollobás, Christian Borgs, Jennifer Chayes, and Oliver Riordan. Directed scale-free graphs. In *Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 132–139. Society for Industrial and Applied Mathematics, 2003.

[6] P Oscar Boykin and Vwani P Roychowdhury. Leveraging social networks to fight spam. *Computer*, 38(4):61–68, 2005.

[7] Jilin Chen, Werner Geyer, Casey Dugan, Michael Muller, and Ido Guy. Make new friends, but keep the old: recommending people on social networking sites. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 201–210. ACM, 2009.

[8] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

[9] Don Coppersmith and Shmuel Winograd. Matrix multiplication via arithmetic progressions. In *Proceedings of the nineteenth annual ACM symposium on Theory of computing*, pages 1–6. ACM, 1987.

[10] Richard Durrett. *Random graph dynamics*, volume 200. Cambridge university press Cambridge, 2007.

[11] Giorgio Fagiolo. Clustering in complex directed networks. *Physical Review E*, 76(2):026107, 2007.

[12] Emilio Ferrara, Roberto Interdonato, and Andrea Tagarelli. Online popularity and topical interests through the lens of instagram. In *Proceedings of the 25th ACM conference on Hypertext and social media*, pages 24–34. ACM, 2014.

[13] Maksym Gabielkov and Arnaud Legout. The complete picture of the twitter social graph. In *Proceedings of the 2012 ACM conference on CoNEXT student workshop*, pages 19–20. ACM, 2012.

[14] Maksym Gabielkov, Ashwin Rao, and Arnaud Legout. Studying social networks at scale: macroscopic anatomy of the twitter social graph. In *ACM SIGMETRICS Performance Evaluation Review*, volume 42, pages 277–288. ACM, 2014.

[15] Mark S Granovetter. The strength of weak ties. In *Social networks*, pages 347–367. Elsevier, 1977.

[16] Beom Jun Kim. Performance of networks of artificial neurons: The role of clustering. *Physical Review E*, 69(4):045101, 2004.

[17] Mihail N Kolountzakis, Gary L Miller, Richard Peng, and Charalampos E Tsourakakis. Efficient triangle counting in large graphs via degree-based vertex partitioning. *Internet Mathematics*, 8(1-2):161–185, 2012.

[18] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM, 2010.

[19] Matthieu Latapy. Main-memory triangle computations for very large (sparse (power-law)) graphs. *Theoretical Computer Science*, 407:458–473, 2008.

[20] François Le Gall. Powers of tensors and fast matrix multiplication. In *Proceedings of the 39th international symposium on symbolic and algebraic computation*, pages 296–303. ACM, 2014.

[21] Jure Leskovec and Eric Horvitz. Planetary-scale views on a large instant-messaging network. In *Proceedings of the 17th international conference on World Wide Web*, pages 915–924. ACM, 2008.

[22] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 177–187. ACM, 2005.

[23] Jure Leskovec, Kevin J Lang, Anirban Dasgupta, and Michael W Mahoney. Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. *Internet Mathematics*, 6(1):29–123, 2009.

[24] Linyuan Lü and Tao Zhou. Link prediction in complex networks: A survey. *Physica A: statistical mechanics and its applications*, 390(6):1150–1170, 2011.

[25] Alan Mislove, Hema Swetha Koppula, Krishna P. Gummadi, Peter Druschel, and Bobby Bhattacharjee. Growth of the flickr social network. In *Proceedings of the 1st ACM SIGCOMM Workshop on Social Networks (WOSN'08)*, August 2008.

[26] Alan Mislove, Massimiliano Marcon, Krishna P Gummadi, Peter Druschel, and Bobby Bhattacharjee. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 29–42. ACM, 2007.

[27] Seth A Myers, Aneesh Sharma, Pankaj Gupta, and Jimmy Lin. Information network or social network?: the structure of the twitter follow graph. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 493–498. ACM, 2014.

[28] Mark EJ Newman. Ego-centered networks and the ripple effect. *Social Networks*, 25(1):83–95, 2003.

[29] Tore Opsahl and Pietro Panzarasa. Clustering in weighted networks. *Social networks*, 31(2):155–163, 2009.

[30] Ryan Rossi and Nesreen Ahmed. The network data repository with interactive graph analytics and visualization. In *AAAI*, volume 15, pages 4292–4293, 2015.

[31] Seyed-Vahid Sanei-Mehri, Ahmet Erdem Sariyuce, and Srikanta Tirthapura. Butterfly counting in bipartite networks. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2150–2159. ACM, 2018.

[32] Jari Saramäki, Mikko Kivelä, Jukka-Pekka Onnela, Kimmo Kaski, and Janos Kertesz. Generalizations of the clustering coefficient to weighted complex networks. *Physical Review E*, 75(2):027105, 2007.

[33] Thomas Schank and Dorothea Wagner. Finding, counting and listing all triangles in large graphs, an experimental study. In *International workshop on experimental and efficient algorithms*, pages 606–609. Springer, 2005.

[34] Hamza Shaban. https://www.washingtonpost.com/technology/2019/02/07/twitter-reveals-its-daily-active-user-numbers-first-time/.

[35] Nitai B Silva, Ren Tsang, George DC Cavalcanti, and Jyh Tsang. A graph-based friend recommendation system using genetic algorithm. In *Evolutionary Computation (CEC), 2010 IEEE Congress on*, pages 1–7. IEEE, 2010.

[36] Steven H Strogatz. Exploring complex networks. *nature*, 410(6825):268, 2001.

[37] Johan Ugander, Brian Karrer, Lars Backstrom, and Cameron Marlow. The anatomy of the facebook social graph. *arXiv preprint arXiv:1111.4503*, 2011.

[38] Alexei Vázquez. Growing network with local rules: Preferential attachment, clustering hierarchy, and degree correlations. *Physical Review E*, 67(5):056104, 2003.

[39] Jia Wang, Ada Wai-Chee Fu, and James Cheng. Rectangle counting in large bipartite graphs. In *2014 IEEE International Congress on Big Data*, pages 17–24. IEEE, 2014.

[40] Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world'networks. *nature*, 393(6684):440, 1998.

Chapter 4

A Random Growth Model with any Real or Theoretical Degree Distribution

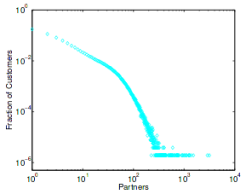
4.1 Introduction

Complex networks appear in the empirical study of real world networks from various domains, such that social, biology, economy, technology, ... Most of those networks exhibit common properties, such as high clustering coefficient, communities, ... Probably the most studied of those properties is the degree distribution (named DD in the rest of the Chapter), which is often observed as following a power-law distribution. Random network models have thus focused on being able to build graphs exhibiting power-law DDs, such as the well-known Barabasi-Albert model [2] or the Chung-Lu model [7], but also models for directed networks [4] or for networks with communities [24]. However, this is common to find real networks with DDs not perfectly following a power-law. For instance for social networks, Facebook has been shown to follow a broken power-law[1] [13], while Twitter only has the distribution tail following a power-law and some atypical behaviors due to Twitter's policies, as we report in Section 4.5.1.

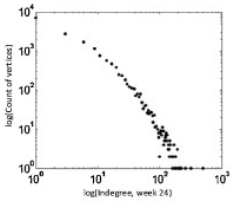
It is yet crucial to build models able to reproduce the properties of real networks. Indeed, some studies such as fake news propagation or evolution over time of the networks cannot always be done empirically, for technical or ethical reasons. Carrying out simulations with random networks created with well-built models is a solution to study real networks without directly experimenting on them. Those models have to create networks with similar properties as real ones, while staying as simple as possible.

In this Chapter, we propose a random growth model able to create graphs with almost any (under some conditions) given DD. Classical models usually choose the nodes receiving new edges proportionally to a linear attachment function $f(i) = i$ (or $f(i) = i + b$) [2, 4]. The theoretical DD of the networks generated by those models is computed using a recurrence equation. The main idea of this Chapter is to reverse this recurrence equation to express the attachment function f as a function of the DD. This way, for a given DD, we can compute the associated attachment function,

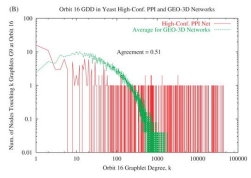
[1] We call a broken power-law a concatenation of two power-laws, as defined in [15].



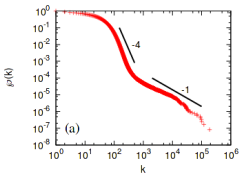
(a) DD of the number of unique callers and callees from a mobile phone operator. [25]



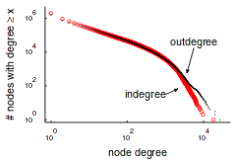
(b) In-DD between shop-to-shop recommendations from an online marketplace. [26]



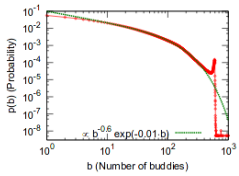
(c) Graphlet DD from a biological model. [22]



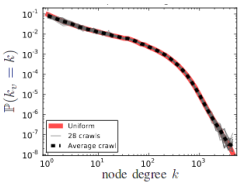
(d) DD of users of Cyworld, the largest online social network of South Korea. [1]



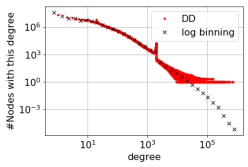
(e) DDs of users of Flickr, an online social network. [6]



(f) DD of the length of the contact list in Microsoft Messenger network. [16]



(g) DD of the number of friends from FaceBook, a social network. [13]



(h) Out-DD of the number of followees on Twitter. [27]

Figure 4.1: DDs extracted from different seminal papers studying networks from various domains.

and use it in a proposed random growth model to create graphs with the wanted DD. The given DD can either be theoretical, or extracted from a real network.

We compute the attachment function associated with some classical DD, homogeneous ones such as the geometric distribution, and heterogeneous ones such as exact power-law and broken power-law. We also study the undirected DD of a Twitter snapshot of 505 million nodes and 23 billion edges, extracted by Gabielkov et al. [11] and made available by the authors. We notice it has an atypical shape, due to Twitter's policies. We compute empirically the associated attachment function, and use the model to build random graphs with this DD. A necessary condition is that the given DD must be defined for all degrees under the (arbitrary chosen) maximum value. However this condition can be circumvented doing an interpolation between existing points to estimate the missing ones, as discussed in Section 4.5. Finally, we study some connections between attachment functions and probability distributions in Section 4.6. More precisely, we show that in our model, unless for some really unusual cases, the probability distribution is heavy-tailed if and only if the attachment function diverges.

The rest of the Chapter is organized as follows. We first discuss the related work in Section 4.2. In Section 4.3, we present the new model, and invert the recurrence equation to find the relation between the attachment function and the DD. We apply this relation to compute the attachment function associated to some theoretical distributions in Section 4.4. In Section 4.5 we apply our model on a real-world DD, the undirected DD of Twitter. We finally show the link between the divergence of the attachment function and the heavy-tailed property of the probability distribution in Section 4.6.

4.2 Related Work

The degree distribution has been computed for a lot of networks, in particular for social networks such as Facebook [13] or Microsoft Messenger [16]. Note that Myers et al. have also studied DDs for Twitter in [19], using a different dataset than the one of [11].

Questioning the relevance of power-law fits is not new: for instance, Clauset et al. [9] or Lima-Mendez and van Helden [17] have already deeply questioned the myth of power-law -as Lima-Mendez and van Helden call it-, and develop tools to verify if a distribution can be considered as a power-law or not. Clauset et al. apply the developed tools on 24 distributions extracted from various domains of literature, which have all been considered to be power-laws. Among them, "17 of the 24 data sets are consistent with a power-law distribution", and "there is only one case in which the power law appears to be truly convincing, in the sense that it is an excellent fit to the data and none of the alternatives carries any weight". In the continuity of this work, Broido and Clauset study in [5] the DD of nearly 1000 networks from various domains, and conclude that "fewer than 36 networks (4%) exhibit the strongest level of evidence for scale-free structure".

The study of Clauset et al. [9] only considered distributions which have a power-law shape when looking at the distribution in log-log. As a complement, we gathered DDs from literature which clearly do not follow power-law distributions to show their diversity. We extracted from literature DDs of networks from various domains: biology, economy, computer science, ... Each presented DD comes from a seminal well cited paper of the respective domains. They are gathered in Figure 4.1. Various shapes can be observed from those DDs, which could (by eyes) be associated with exponential (Fig. 4.1b, 4.1c), broken power-law (Fig. 4.1a, 4.1e, 4.1g), or even some kind of inverted broken power-law (Fig 4.1d). We also observe DDs with specific behaviors (Fig. 4.1f, 4.1h).

The first proposed models of random networks, such as the Erdős–Rényi model [10], build networks with a homogeneous DD. The observation that a lot of real-world networks follow power-law DDs lead Albert and Barabasi to propose their famous model with linear preferential attachment [2]. It has been followed by a lot of random growth models, e.g. [4, 7] also giving a DD in power-law. A few models permit to build networks with any DD: for instance, the configuration model [3, 20] takes as parameter a DD P and a number of nodes n, creates n nodes with a degree randomly picked following P, then randomly connects the half-edges of every node. Goshal and Newman propose in [12] a model generating non-growing networks (where, at each time-step, a node is added and another is deleted) which can achieve any DD, using a method close to the one proposed in this Chapter. However, both of those models generate non-growing networks, while most real-world networks are constantly growing.

4.3 Presentation of the model

The proposed model is a generalization of the model introduced by Chung and Lu in [7]. At each time step, we have either a node event or an edge event. During a node event, a node is added with an edge attached to it; during an edge event, an edge is added between two existing nodes. Each node to which the edge is

connected is randomly chosen among all nodes with a probability proportional to a given function f, called the *attachment function*. The model is as follows:

- ▷ We start with an initial graph G_0.
- ▷ At each time step t:
  - With probability p: we add a node u, and an edge (u, v) where the node v is randomly chosen among all existing nodes with a probability $\frac{f(deg(v))}{\sum_{w \in V} f(deg(w))}$;
  - With probability $(1-p)$: we add an edge (u, v) where the nodes u and v are randomly chosen among all existing nodes with a probability $\frac{f(deg(u))}{\sum_{w \in V} f(deg(w))}$ and $\frac{f(deg(v))}{\sum_{w \in V} f(deg(w))}$.

Note that the Chung-Lu model is the particular case for which $f(i) = i$ for all $i \geq 1$. We call *generalized Chung-Lu model* the proposed model where $f(i) = i + b$, for all $i \geq 1$ with $b > -1$.

4.3.1 Connection between the attachment function and the degree distribution

The common way to find the DD of classical random growth models is to study the recurrence equation of the evolution of the number of nodes with degree i between two time steps. This equation can sometimes be easily solved, sometimes not. But what matters for us is that the common process is to start from a given model -thus an attachment function f-, and use the recurrence equation to find the DD P. In this section, we show that the recurrence equation of the proposed model can be reversed such that, if P if given, we can find an associated attachment function f.

Theorem 2. *Let P be a probability distribution of finite mean μ and such that the following function h is bounded:*

$$h(i) = \frac{P(k > i+1)}{P(i+1)} - \frac{P(k > i)}{P(i)}.$$

In the proposed model, if p is chosen as $p = \frac{1}{\mu}$ and if the attachment function is chosen as:

$$\forall i \geq 1, f(i) = \frac{1}{P(i)} \sum_{k=i+1}^{\infty} P(k), \tag{4.1}$$

then the DD of the created graph is distributed according to P.

Remark 2. *The condition on p comes from the fact that, by construction of the model, we have $\mathbb{E}[N(t)] = pt$ and $\mathbb{E}(|E|(t)) = t$ with $|E|(t)$ the number of edges at time t. This leads to a mean-degree of $\frac{1}{p}$.*

Remark 3. *Note that Equation 2 can also be expressed as $f(i) = \frac{P(k>i)}{P(i)}$.*

For a given probability law, Theorem 2 can be used to compute the attachment function which, when used in the model, will give this probability law as DD. The remainder of this subsection is intented to prove Theorem 2.

Proof of Theorem 2

Let $N(i,t)$ be the random variable corresponding to the number of nodes of degree i at time t in the graph, $N(t)$ the total number of nodes at time t, and $P(i) = \lim\limits_{t\to+\infty} \mathbb{E}[\frac{N(i,t)}{N(t)}]$ the probability that a random node has degree i in the asymptotic DD.

Before proving Theorem 2, we need some results on the concentration of $N(t)$ and $\sum_{j\geq 1} f(j)N(j,t)$. We start with $N(t)$. We will need the following lemma:

Lemma 4 (Chernoff bounds, consult Chapter 4.2 in [18])**.** *Let $X_1, X_2, \ldots, X_t$ be independent indicator random variables with $\mathbb{P}[X_i = 1] = p_i$ and $\mathbb{P}[X_i = 0] = 1 - p_i$. Let $X = \sum_{i=1}^{t} X_i$ and $\mu = \mathbb{E}[X] = \sum_{i=1}^{t} p_i$. Then*

$$\mathbb{P}[|X - \mu| \geqslant \delta\mu] \leq 2e^{-\mu\delta^2/3}.$$

$N(t)$ is a random variable following a binomial distribution $N(t) \sim B(t,p) + n_0$, with n_0 the number of nodes in the initial graph. We can thus use Lemma 11 on $N(t)$; since $\mathbb{E}[N(t)] = pt$, setting $\delta = \sqrt{\frac{9\ln t}{pt}}$ we get:

Corollary 1.

$$\mathbb{P}[|N(t) - pt| \geqslant \sqrt{9pt\ln t}] \leq 2/t^3. \tag{4.2}$$

We also have the following result on P:

Lemma 5. $P(i) \underset{t\to+\infty}{\sim} \frac{\mathbb{E}[N(i,t)]}{pt}$

Proof. For more clarity in this proof let us denote $N(t)$ as N_t and $N(i,t)$ as $N_{i,t}$. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be the probability space on which random variables $N_{i,t}$ and N_t are defined. Thus $N_{i,t} : \Omega \to \mathbb{R}$ and $N_t : \Omega \to \mathbb{R}$. Let $\Omega_1 \subseteq \Omega$ denote the set of all $\omega \in \Omega$ such that $N_t(\omega) \in (\mathbb{E}[N_t] - \sqrt{9pt\ln t}, \mathbb{E}[N_t] + \sqrt{9pt\ln t})$. By Corollary 2 we know that $\sum_{\omega\in\Omega\setminus\Omega_1} \mathbb{P}[\omega] \leqslant 2/t^3$. Using the fact that $\mathbb{E}[N_t] = pt$ and that for each ω $\frac{N_{i,t}(\omega)}{N_t(\omega)} \leqslant 1$ we get

$$\mathbb{E}\left[\frac{N_{i,t}}{N_t}\right] = \sum_{\omega\in\Omega} \frac{N_{i,t}(\omega)}{N_t(\omega)}\mathbb{P}[\omega] = \sum_{\omega\in\Omega_1} \frac{N_{i,t}(\omega)}{N_t(\omega)}\mathbb{P}[\omega] + \sum_{\omega\in\Omega\setminus\Omega_1} \frac{N_{i,t}(\omega)}{N_t(\omega)}\mathbb{P}[\omega] \tag{4.3}$$

$$\leqslant \sum_{\omega\in\Omega} \frac{N_{i,t}(\omega)}{\mathbb{E}[N_t] - \sqrt{9pt\ln t}}\mathbb{P}[\omega] + \sum_{\omega\in\Omega\setminus\Omega_1} 1\cdot\mathbb{P}[\omega] \tag{4.4}$$

$$\leqslant \frac{\mathbb{E}[N_{i,t}]}{\mathbb{E}[N_t] - \sqrt{9pt\ln t}} + 2/t^3 \sim \frac{\mathbb{E}[N_{i,t}]}{pt}. \tag{4.5}$$

On the other hand, since $N_{i,t} \leq t$,

$$\mathbb{E}\left[\frac{N_{i,t}}{N_t}\right] \geqslant \sum_{\omega\in\Omega_1} \frac{N_{i,t}(\omega)}{N_t(\omega)}\mathbb{P}[\omega] \geqslant \sum_{\omega\in\Omega_1} \frac{N_{i,t}(\omega)}{\mathbb{E}[N_t] + \sqrt{9pt\ln t}}\mathbb{P}[\omega] \tag{4.6}$$

$$= \frac{1}{\mathbb{E}[N_t] + \sqrt{9pt\ln t}} \left(\mathbb{E}[N_{i,t}] - \sum_{\omega\in\Omega\setminus\Omega_1} N_{i,t}(\omega)\mathbb{P}[\omega]\right) \tag{4.7}$$

$$\geqslant \frac{1}{\mathbb{E}[N_t] + \sqrt{9pt\ln t}} \left(\mathbb{E}[N_{i,t}] - \sum_{\omega\in\Omega\setminus\Omega_1} t\cdot\mathbb{P}[\omega]\right) \tag{4.8}$$

$$\geqslant \frac{\mathbb{E}[N_{i,t}]}{\mathbb{E}[N_t] + \sqrt{9pt\ln t}} - \frac{t\cdot 2/t^3}{\mathbb{E}[N_t] + \sqrt{9pt\ln t}} \tag{4.9}$$

$$\sim \frac{\mathbb{E}[N_{i,t}]}{pt}. \tag{4.10}$$

□

We now discuss the concentration of $\sum_{j\geq 1} f(j)N(j,t)$. Let us define

$$Z_t = \sum_{j\geq 1} f(j)N(j,t). \tag{4.11}$$

Using the following lemma from [14]:

Lemma 6 (Hoeffding's inequality, [14])**.** *Let $X_1, X_2, \ldots, X_t$ be independent random variables such that $\mathbb{P}[X_k \in [a_k, b_k]] = 1$. Let $X = \sum_{k=1}^{t} X_k$. Then*

$$\mathbb{P}[|X - \mathbb{E}[X]| \geqslant \delta] \leqslant 2\exp\left\{-\frac{2\delta^2}{\sum_{k=1}^{t}(a_k - b_k)^2}\right\}. \tag{4.12}$$

We can show that:

Lemma 7. *If the following condition is satisfied:*

$$\exists K/\forall i \geqslant 1, |f(i+1) - f(i)| \leqslant K \tag{4.13}$$

Then:

$$\mathbb{P}[|Z_t - \mathbb{E}[Z_t]| \geqslant \sqrt{32K^2t\ln t}] = \mathcal{O}\left(\frac{1}{t^4}\right). \tag{4.14}$$

Proof. First, remind that Z_t can either be express as $Z_t = \sum_{j\geq 1} f(j)N(j,t)$ or $Z_t = \sum_{u\in V_t} f(deg_t(u))$, with $deg_t(u)$ the degree of node u at time t. But Z_t can also be express as a sum of independent random variables $X_1 + X_2 + ... + X_t$, with X_k the variation of Z_k during the time step k, i.e. $X_k = Z_k - Z_{k-1}$. In practice, X_k can take those different values:

- With probability p, a node and an edge are added to the graph, and $X_k = f(deg_k(u) + 1) - f(deg_k(u)) + f(1)$, with u the chosen node at time step k;
- With probability $(1-p)$, an edge is added between two existing nodes, and $X_k = f(deg_k(u) + 1) - f(deg_k(u)) + f(deg_k(v) + 1) - f(deg_k(v))$, with u and v the chosen nodes.

Using the condition on f, we see that we can bound X_k by $-2K \leqslant X_k \leqslant 2K$. We can thus apply Lemma 13 with $X = \sum_{k=1}^{t} X_k = Z_t$, $a_i = -2K$ and $b_i = 2K$ to obtain:

$$\mathbb{P}[|Z_t - \mathbb{E}[Z_t]| \geqslant \delta] \leqslant 2\exp\left\{-\frac{2\delta^2}{t(4K)^2}\right\}. \tag{4.15}$$

Now, setting $\delta = \sqrt{32K^2 t \ln t}$ we get:

$$\mathbb{P}[|Z_t - \mathbb{E}[Z_t]| \geqslant \sqrt{32K^2 t \ln t}] \leqslant 2\exp\left\{-\frac{2 \cdot 32K^2 t \ln t}{t(4K)^2}\right\} = \mathcal{O}\left(\frac{1}{t^4}\right). \tag{4.16}$$

□

We will finally need the following lemma from [8]:

Lemma 8 (Compare Chapter 3.3 in [8])**.** *Let (a_t), (b_t), (c_t) be three sequences such that $a_{t+1} = (1 - \frac{b_t}{t})a_t + c_t$, $\lim\limits_{t\to+\infty} b_t = b > 0$, and $\lim\limits_{t\to+\infty} c_t = c$. Then $\lim\limits_{t\to+\infty} \frac{a_t}{t}$ exists and equals $\frac{c}{1+b}$.*

We are now ready to prove Theorem 2.

Proof of Theorem 2. During the proof, we will consider the following conditions as true:

C1) $\exists K / \forall i \geqslant 1, |f(i+1) - f(i)| \leqslant K$,

C2) $\sum_{j\geq 1} f(j)P(j) = \mu$, $\mu \in \mathbb{R}_+^*$.

where we remind that P is defined as $P(i) = \lim\limits_{t\to+\infty} \mathbb{E}[\frac{N(i,t)}{N(t)}]$. We will verify those conditions are indeed satisfied for the chosen f at the end of the proof. We will verify at the end of the proof that the first condition is equivalent to the condition of Theorem 2, and the second condition is indeed satisfied for the chosen f.

We consider the variation of the number of nodes of degree i $N(i,t)$ between a time step from t to (t+1). During this time step, a node with degree $i-1$ may gain a degree and thus increases by 1 the number of nodes of degree i. This happens with a probability $p+2(1-p)$ (the mean number of half-edges connected to existing nodes during a time step) $\times \frac{f(i-1)}{\sum_{j\geq 1} f(j)N(j,t)}$ (the probability for this particular node of degree $i-1$ to be chosen). Since it is the same for all nodes of degree $i-1$, the number of nodes going from degree $i-1$ to i during a time step is $\big(p + 2(1-p)\big) \times \frac{f(i-1)}{\sum_{j\geq 1} f(j)N(j,t)} \times N(i-1,t)$. In the same way, a node with degree i may be connected to an edge, thus becoming a node with degree $i+1$ and decreasing the number of nodes of degree i. Finally, with probability p, a node of degree 1 is added. Gathering those contributions give the following equation:

$$\begin{aligned} &N(i,t+1) - N(i,t) = \\ &p\delta_{i,1} + (2-p)\frac{f(i-1)}{\sum\limits_{j\geq 1} f(j)N(j,t)}N(i-1,t) - (2-p)\frac{f(i)}{\sum\limits_{j\geq 1} f(j)N(j,t)}N(i,t) \end{aligned} \tag{4.17}$$

where $\delta_{i,j}$ is the Kronecker delta. The first term of the right hand is the probability of addition of a node. The second (resp. third) term is the probability that a

node of degree $i-1$ (resp. i) gets chosen to be the end of an edge. The factor $(2-p) = p + 2(1-p)$ comes from the fact that this happens with probability p during a node event (connection of a single half-edge) and with probability $2(1-p)$ during an edge event (possible connection of 2 half-edges).

We take the expectation on both sides and use Lemma 7 to obtain:

$$\mathbb{E}[N(i,t+1)] - \mathbb{E}[N(i,t)] = p\delta_{i,1} \tag{4.18}$$

$$+ (2-p)\frac{f(i-1)}{\sum\limits_{j\geq 1} f(j)\mathbb{E}[N(j,t)] + \mathcal{O}\left(\sqrt{t\ln t}\right)}\mathbb{E}[N(i-1,t)] \tag{4.19}$$

$$- (2-p)\frac{f(i)}{\sum\limits_{j\geq 1} f(j)\mathbb{E}[N(j,t)] + \mathcal{O}\left(\sqrt{t\ln t}\right)}\mathbb{E}[N(i,t)]$$

We denote $g(i) = \frac{2-p}{p}\frac{f(i)}{\sum_{j\geq 1} f(j)P(j)}$. We first show that $g(i) = \frac{1}{P(i)}\sum\limits_{k=i+1}^{\infty} P(k)$. We will then show that we can choose $f = g$.
For $i = 1$, Equation 4.18 becomes:

$$\mathbb{E}[N(1,t+1)] - \mathbb{E}[N(1,t)] = p - (2-p)\frac{f(1)}{\sum\limits_{j\geq 1} f(j)\mathbb{E}[N(j,t)] + \mathcal{O}\left(\sqrt{t\ln t}\right)}\mathbb{E}[N(1,t)]. \tag{4.20}$$

Taking:

$$a_t = \frac{\mathbb{E}[N(1,t)]}{p},$$

$$b_t = \frac{(2-p)f(1)}{p\sum_{j\geq 1} f(j)\frac{\mathbb{E}[N(j,t)]}{pt} + \mathcal{O}\left(\sqrt{\frac{\ln t}{t}}\right)},$$

$$c_t = 1,$$

we have $\lim\limits_{t\to+\infty} b_t = g(1) > 0$ and $\lim\limits_{t\to+\infty} c_t = 1$. We can thus apply Lemma 8 (and use Lemma 5 to recognize $P(1)$):

$$\lim_{t\to+\infty}\frac{\mathbb{E}[N(1,t)]}{pt} = P(1) = \frac{1}{1+g(1)}, \tag{4.21}$$

Now, $\forall i \geq 2$, taking:

$$a_t = \frac{\mathbb{E}[N(i,t)]}{p},$$

$$b_t = \frac{(2-p)f(i)}{p\sum_{j\geq 1} f(j)\frac{\mathbb{E}[N(j,t)]}{pt} + \mathcal{O}\left(\sqrt{\frac{\ln t}{t}}\right)},$$

$$c_t = \frac{(2-p)f(i-1)}{p\sum_{j\geq 1} f(j)\frac{\mathbb{E}[N(j,t)]}{pt} + \mathcal{O}\left(\sqrt{\frac{\ln t}{t}}\right)}\frac{\mathbb{E}[N(i-1,t)]}{pt},$$

we have $\lim\limits_{t\to+\infty} b_t = g(i) > 0$ and $\lim\limits_{t\to+\infty} c_t = g(i-1)P(i-1)$. Lemma 8 and Lemma 5 give:

$$\lim_{t\to+\infty}\frac{\mathbb{E}[N(i,t)]}{pt} = P(i) = \frac{g(i-1)P(i-1)}{1+g(i)}. \tag{4.22}$$

Name	P(i)	f(i)	Condition
Generalized Chung-Lu	$C\frac{\Gamma(i+b)}{\Gamma(i+b+\alpha)}$	$\frac{1}{\alpha-1}i+\frac{b}{\alpha-1}$	$p=\frac{\alpha-2}{\alpha+b-1}$
Exact Power-Law	$\frac{i^{-\alpha}}{\zeta(\alpha)}$	$\frac{\zeta(\alpha,i+1)}{i^{-\alpha}}$	$p=\frac{\zeta(\alpha)}{\zeta(\alpha-1)}$
Geometric Law	$q(1-q)^{i-1}$	$\frac{1-q}{q}$	$p=q$
Broken Power-Law	$\begin{cases} C\frac{\Gamma(i+b_1)}{\Gamma(i+b_1+\alpha_1)} & \text{if } i\leq d \\ C\gamma\frac{\Gamma(i+b_2)}{\Gamma(i+b_2+\alpha_2)} & \text{if } i>d \end{cases}$	cf. eq. 4.36& 4.37	cf. eq. 4.35

Table 4.1: Attachment functions f and conditions on p for some classical probability distributions P. $\zeta(s)$ is the Riemann zeta function, $\zeta(s,q)$ the Hurwitz zeta function.

Iterating over Equation 4.22, we express g as a function of P:

$$\begin{aligned} g(i)P(i) &= g(i-1)P(i-1) - P(i) \\ &= g(i-2)P(i-2) - P(i-1) - P(i) \\ &= \cdots \\ &= g(1)P(1) - \sum_{k=2}^{i} P(k) \\ &= 1 - \sum_{k=1}^{i} P(k) \end{aligned}$$

$$\implies g(i) = \frac{1}{P(i)} \sum_{k=i+1}^{\infty} P(k) \tag{4.23}$$

where we used Equation 4.21 to replace $g(1)P(1)$.
Now, notice that:

$$\sum_{k=1}^{\infty} g(k)P(k) = \sum_{k=1}^{\infty} \frac{2-p}{p} \frac{f(k)}{\sum_{k'=1}^{\infty} f(k')P(k')} P(k) = \frac{(2-p)}{p}. \tag{4.24}$$

So $g(i)$ satisfies $g(i) = \frac{2-p}{p}\frac{g(i)}{\sum_{k=1}^{\infty} g(k)P(k)}$. Hence the attachment function can be chosen as $f = g$.
We finally have to verify the conditions we put at the beginning of the proof are true. The first condition is equivalent to the condition of Theorem 2 for the given f. The second condition is given by Equation 4.24, which conclude the proof. □

4.4 Application to some distributions

We now apply Equation 4.1 to compute the attachment function for some classical distributions. We first start in Section 4.4.1 from the distribution obtained with the generalized Chung-Lu model to show we find a linear dependence, as expected. We then compute in Section 4.4.2 the associated attachment function of the broken power-law distribution. We finally compute the exact power-law and geometric law distributions in Sections 4.4.3 and 4.4.4. Table 4.1 summarizes those results.

4.4.1 Preliminary: Generalized Chung-Lu model

As a first example, by taking a power-law DD, we should be able to find a linear probability distribution for the generalized Chung-Lu model.

In the general Chung-Lu model, we can show that the real DD is not an exact power-law but a fraction of Gamma function -equivalent to a power-law for high degrees- of the form:

$$\forall i \geq 1, P(i) = C\frac{\Gamma(i+b)}{\Gamma(i+b+\alpha)} \underset{i\gg 1}{\sim} i^{-\alpha} \tag{4.25}$$

where $C = (\alpha-1)\frac{\Gamma(b+\alpha)}{\Gamma(b+1)}$, and $\alpha > 2$. The choice of α determines the slope of the DD, while the choice of b determines the mean-degree of the graph.

Expression of p: The condition on p from Theorem 4.1 gives:

$$\begin{aligned}\frac{1}{p} = \sum_{k=1}^{\infty} kP(k) &= (\alpha-1)\frac{\Gamma(b+\alpha)}{\Gamma(b+1)} \times \frac{\alpha^2+\alpha(2b-1)+b(b-1)}{(\alpha-2)(\alpha-1)}\frac{\Gamma(b+1)}{\Gamma(\alpha+b+1)} \\ \implies p &= \frac{(\alpha-2)}{\alpha+b-1}\end{aligned} \tag{4.26}$$

Condition of Theorem 2: We first verify the necessary condition on P of Theorem 2. Since $P(X > i) = \frac{C}{\alpha-1}\frac{\Gamma(i+b)}{\Gamma(i+b+\alpha-1)}$, we have:

$$h(i) = \frac{P(k>i+1)}{P(i+1)} - \frac{P(k>i)}{P(i)} \tag{4.27}$$

$$= \frac{i+b+\alpha}{\alpha-1} - \frac{i+b+\alpha-1}{\alpha-1} \tag{4.28}$$

$$= \frac{1}{\alpha-1} \tag{4.29}$$

Thus $g(i)$ is constant and verify the condition. Note that this result is expected since we know that, in this classical case, f is linear and so $f(i+1)-f(i)$ is indeed expected to be constant.

Attachment function f: Using Theorem 2:

$$f(i) = \frac{1}{P(i)}\sum_{k\geq i+1} P(k) = \frac{\Gamma(i+b+\alpha)}{\Gamma(i+b)}\frac{\Gamma(i+b+1)}{(\alpha-1)\Gamma(i+\alpha+b)} \tag{4.30}$$

$$\implies f(i) = \frac{1}{\alpha-1}i + \frac{b}{\alpha-1} \tag{4.31}$$

As expected, we find a linear attachment function. To create a graph with a wanted slope α and mean-degree p^{-1}, one only has to choose α as the wanted slope and b following equation 4.26. In the particular case $b = 0$, we recover the Chung-Lu model of [7], with a slope of $\alpha = 2 + \frac{p}{2-p}$ as expected.

4.4.2 Broken Power-law

We now study the case of a broken power-law, corresponding to the DD of real world complex networks, as discussed in Section 4.2. which was the one we were interested in initially. We consider a distribution of the form:

$$P(i) = \begin{cases} C\frac{\Gamma(i+b_1)}{\Gamma(i+b_1+\alpha_1)} & \text{if } i \leq d \\ C\gamma\frac{\Gamma(i+b_2)}{\Gamma(i+b_2+\alpha_2)} & \text{if } i > d \end{cases} \tag{4.32}$$

where d, b_1, α_1, b_2, and α_2 are parameters of our distribution such that $\alpha_1 > 2$, $\alpha_2 > 2$, C a normalisation constant, and γ chosen in order to obtain continuity for $i = d$. As seen in section 4.4.1, the ratio of gamma functions is close to a power-law as soon as i gets large. Hence, this distribution corresponds to two powers-laws, with different slopes, and a switch between the two at the value d.

We can easily find the continuity constant γ, since it verifies:

$$\frac{\Gamma(d+b_1)}{\Gamma(d+b_1+\alpha_1)} = \gamma\frac{\Gamma(d+b_2)}{\Gamma(d+b_2+\alpha_2)} \implies \gamma = \frac{\Gamma(d+b_1)\Gamma(d+b_2+\alpha_2)}{\Gamma(d+b_1+\alpha_1)\Gamma(d+b_2)}. \tag{4.33}$$

Constraints on C and p: The value of C can be computed by summing over all degrees:

$$C = \Big(\sum_{k=1}^{\infty} P(k)\Big)^{-1} = \Big(\frac{1}{\alpha_1 - 1}\frac{\Gamma(b_1+1)}{\Gamma(\alpha_1+b_1)} + \frac{\Gamma(b_1+d)}{\Gamma(\alpha_1+b_1+d)}\big(\frac{b_2+d}{\alpha_2-1} - \frac{b_1+d}{\alpha_1-1}\big)\Big)^{-1} \tag{4.34}$$

Using the condition in Theorem 4.1, p is defined by the following equation:

$$\begin{aligned} \frac{1}{pC} &= \sum_{k=1}^{d} k\frac{\Gamma(k+b_1)}{\Gamma(k+b_1+\alpha_1)} + \gamma\sum_{k=d+1}^{\infty} k\frac{\Gamma(k+b_2)}{\Gamma(k+b_2+\alpha_2)} \\ &= \frac{\alpha_1^2 + \alpha_1(2b_1-1) + b_1(b_1-1)}{(\alpha_1-2)(\alpha_1-1)}\frac{\Gamma(b_1+1)}{\Gamma(\alpha_1+b_1+1)} \\ &- \frac{\alpha_1^2(d+1) + \alpha_1(b_1(d+2)+d^2-1) + b_1(b_1-1) - d(d+1)}{(\alpha_1-2)(\alpha_1-1)}\frac{\Gamma(b_1+d+1)}{\Gamma(\alpha_1+b_1+d+1)} \\ &+ \gamma\frac{\alpha_2^2(d+1) + \alpha_2(b_2(d+2)+d^2-1) + b_2(b_2-1) - d(d+1)}{(\alpha_2-2)(\alpha_2-1)}\frac{\Gamma(b_2+d+1)}{\Gamma(\alpha_2+b_2+d+1)} \end{aligned} \tag{4.35}$$

Condition of Theorem 2: Let us call $K_d = \max_{i \leq d} g(i)$. From Equation 4.29 we know that, $\forall i > d, h(i) = \frac{1}{\alpha-1}$ since we are in the same case than the Generalized Chung-Lu model. Thus h is bounded by $\max(K_d, \frac{1}{\alpha-1})$.

Attachment function f: For the computation of the attachment function, we have to distinguish two cases:

Case 1: $i \geq d$

$$f(i) = \frac{1}{P(i)} \sum_{k=i+1}^{\infty} P(k) = \frac{\Gamma(i+b_2+\alpha_2)}{\Gamma(i+b_2)} \sum_{k=i+1}^{\infty} \frac{\Gamma(k+b_2)}{\Gamma(k+b_2+\alpha_2)}$$
$$= \frac{\Gamma(i+b_2+\alpha_2)}{\Gamma(i+b_2)} \frac{1}{\alpha_2 - 1} \frac{\Gamma(i+b_2+1)}{\Gamma(i+b_2+\alpha_2)}$$
$$\implies f(i) = \frac{1}{\alpha_2 - 1} i + \frac{b_2}{\alpha_2 - 1} \tag{4.36}$$

We find a linear attachment function: indeed for $i > d$, we only take into account the second power-law, hence we expect to find the same result than in section 4.4.1.

Case 2: $i < d$

$$f(i) = \frac{\Gamma(i+b_1+\alpha_1)}{\Gamma(i+b_1)} \left(\sum_{k=i+1}^{d} \frac{\Gamma(k+b_1)}{\Gamma(k+b_1+\alpha_1)} + \gamma \sum_{k=d+1}^{\infty} \frac{\Gamma(k+b_2)}{\Gamma(k+b_2+\alpha_2)} \right)$$
$$= \frac{\Gamma(i+b_1+\alpha_1)}{\Gamma(i+b_1)} \left(\frac{1}{\alpha_1 - 1} \left(\frac{\Gamma(i+b_1+1)}{\Gamma(i+\alpha_1+b_1)} - \frac{\Gamma(b_1+d+1)}{\Gamma(b_1+\alpha_1+d)} \right) + \frac{\gamma}{\alpha_2 - 1} \frac{\Gamma(b_2+d+1)}{\Gamma(b_2+\alpha_1+d)} \right)$$
$$= \frac{i+b_1}{\alpha_1 - 1} + \frac{\Gamma(i+b_1+\alpha_1)}{\Gamma(i+b_1)} \left(\frac{d+b_2}{\alpha_2 - 1} \frac{\Gamma(b_1+d)}{\Gamma(b_1+\alpha_1+d)} - \frac{1}{\alpha_1 - 1} \frac{\Gamma(b_1+d+1)}{\Gamma(b_1+\alpha_1+d)} \right)$$
$$f(i) = \frac{i+b_1}{\alpha_1 - 1} + \frac{\Gamma(i+b_1+\alpha_1)\Gamma(d+b_1)}{\Gamma(i+b_1)\Gamma(d+b_1+\alpha_1)} \left(\frac{b_2+d}{\alpha_2 - 1} - \frac{b_1+d}{\alpha_1 - 1} \right) \tag{4.37}$$

In this second case, we have a linear part, in addition to a more complicated part. Note that, for $(\alpha_1, b_1) = (\alpha_2, b_2)$, i.e., when the two power-laws are equals, this second term vanishes, letting as expected only the linear part. Figure 4.2a shows the shape of f. We see that, while the second part is linear as discussed before, the first part is sub-linear.

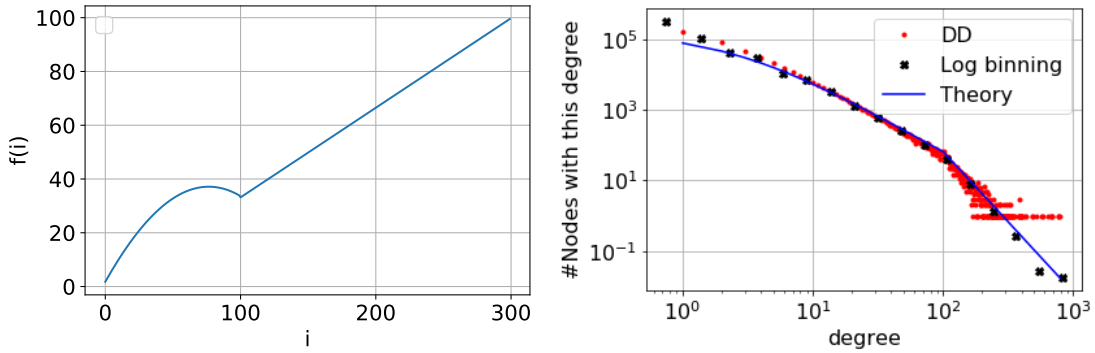
We used this attachment function to build a network using our model. The DD is shown in Figure 4.2b: we see we built a random network with a broken power-law distribution as wanted.

4.4.3 Exact power-law degree distribution

The DD obtained with the Chun-Lu model -and most of other classical models- gives a power-law only for high degrees. We can ask ourselves what would be the attachment function associated with an exact power-law degree distribution of the form $P(i) = \frac{i^{-\alpha}}{\zeta(\alpha)}$, where $\zeta(s) = \sum_{k \geq 1} \frac{1}{k^s}$ is the Riemann zeta function.

Constraints on C and p: We have the following equation for p:

$$\frac{1}{p} = \frac{1}{\zeta(\alpha)} \sum_{k=1}^{\infty} k^{1-\alpha} = \frac{\zeta(\alpha - 1)}{\zeta(\alpha)}$$
$$\implies p = \frac{\zeta(\alpha)}{\zeta(\alpha - 1)}. \tag{4.38}$$



(a) Theoretical attachment function f (b) DD of a random network

Figure 4.2: Theoretical attachment function f and degree distribution of a random network for the broken power-law distribution. Parameters are $N = 5 \cdot 10^5$, $b_1 = b_2 = 1$, $\alpha_1 = 2.1$, $\alpha_2 = 4$ and $d = 100$.

Condition of Theorem 2: To verify the condition, we use the fact, for $\alpha > 1$ and $i > 1$, $\sum_{k \geq i+2} k^{-\alpha} \leq \int_{k \geq i+1} k^{-\alpha} \leq \sum_{k \geq i+1} k^{-\alpha}$. We have:

$$h(i) = \frac{\sum_{k \geq i+2} k^{-\alpha}}{(i+1)^{-\alpha}} - \frac{\sum_{k \geq i+1} k^{-\alpha}}{i^{-\alpha}} \tag{4.39}$$

$$= \int_{k \geq i+1} k^{-\alpha}((i+1)^{\alpha} - i^{\alpha}) \tag{4.40}$$

$$= \frac{1}{\alpha - 1}(i + 1 - (i+1)(\frac{i+1}{i})^{-\alpha}) \tag{4.41}$$

$$= \frac{1}{\alpha - 1}(i + 1 - (i+1)(1 - \frac{\alpha}{i} + o(\frac{1}{i^2})) \tag{4.42}$$

$$= \frac{\alpha}{\alpha - 1} + o(\frac{1}{i}) \tag{4.43}$$

Attachment function: Theorem 2 immediately gives:

$$f(i) = \frac{1}{P(i)} \sum_{k=i+1}^{\infty} P(k) = \frac{\zeta(\alpha, i+1)}{i^{-\alpha}}. \tag{4.44}$$

4.4.4 Geometric law

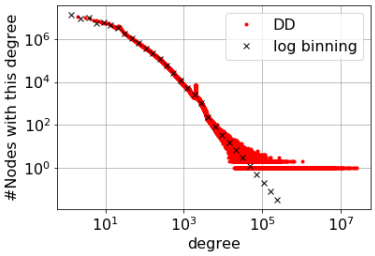
We now study the geometric distribution:

$$\forall i \geq 1, P(i) = q(1-q)^{i-1}. \tag{4.45}$$

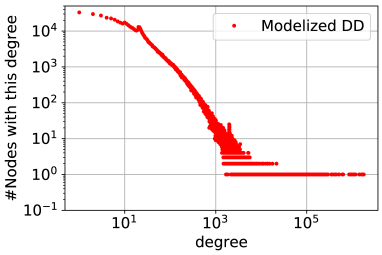
Constraints on p: We have:

$$\frac{1}{p} = \sum_{k \geq 1} kq(1-q)^{k-1} = \frac{q}{(1-q)} \frac{(1-q)}{q^2} = \frac{1}{q} \implies p = q. \tag{4.46}$$

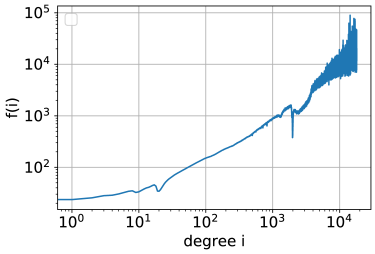
Condition of Theorem 2:



(a) DD of the Twitter's undirected network.



(b) DD of a random network with $8 \cdot 10^5$ nodes using the attachment function of Figure 4.3c.



(c) Attachment function f resulting from the undirected DD of Twitter.

Figure 4.3: Modelization of the undirected Twitter's graph.

$$h(i) = \frac{(1-q)^{i+1}}{q(1-q)^i} - \frac{(1-q)^i}{q(1-q)^{i-1}} = \frac{1-q}{q} - \frac{1-q}{q} = 0 \tag{4.47}$$

Note that this is an expected result since, for the geometric law, the attachment function is constant.

Attachment function: The attachment function is easy to compute:

$$f(i) = \frac{1}{q(1-q)^{i-1}} \sum_{k \geq i+1} q(1-q)^{k-1} = \frac{1}{(1-q)^i} \frac{(1-q)^{i+1}}{q} = \frac{1-q}{q}. \tag{4.48}$$

4.5 Real degree distributions

The model can also be applied to an empirical DD. Indeed, we observe in Theorem 2 that $f(i)$ only depends on the values $P(i)$ which can be arbitrary, that is not following any classical function. This is a good way to model random networks with an atypical DD. As an example, we apply our model on the DD of an undirected version of Twitter, shown as having atypical behavior due to the Twitter policies. We start with a presentation of this DD, then apply our model to build a random graph with this distribution.

4.5.1 Undirected DD of Twitter

For this study, we use a Twitter snapshot from 2012, recovered by Gabielkov and Legout [11] and made available by the authors. This network contains 505 million

nodes and 23 billion edges, making it one of the biggest social graph available nowadays. Each node corresponds to an account, and an arc (u, v) exists if the account u follows the account v. The in- and out-DDs are presented in [27].

In our case, we look at an undirected version of the Twitter snapshot. We consider the degree of each node as being the sum of its in- and out-degrees. The distribution of this undirected graph is presented in Figure 4.3a. We notice two spikes, around $d = 20$ and $d = 2000$. We do not know the reason of the first one (which could be social, or due to recommendation system). The second spike is explained by a specificity of Twitter: until 2015, to avoid bots which were following a very large number of users, Twitter limited the number of possible followings to $\max(2000, \text{number of followers})$. In other words, a user is allowed to follow more than 2000 people only if he is also followed by more than 2000 people. This leads to a lot of accounts with around 2000 followings. This highlights the fact that some networks have their own specificities, sometimes due to intern policies, which cannot be modeled but by a model specifically built for them.

4.5.2 Modelization

Figure 4.3c presents the obtained form of the attachment function f computed using Equation 4.1 with the DD of Twitter. We notice that the overall function is mainly increasing, showing that nodes of higher degrees have a higher chance to connect with new nodes, like in classical preferential attachment models. We also notice two drops, around 20 and 2000. They are associated with the risings on the DD on the same degrees: to increase the amount of nodes with those degrees, the attachment function has to be smaller, so nodes with this degree have less chance to gain new edges.

We finally use our model with the empirical attachment function of Figure 4.3c. Note that, in an empirical study, P can be equal to zero for some degrees, for which no node has this degree in the network. In Twitter, the smallest of those degrees occurs around 18.000. In that case, f cannot be computed. To get around this difficulty, we interpolate the missing values of P, using the two closest smaller and bigger degrees of the missing points. Since we observe the probability distribution on a log-log scale, we interpolate between the two points as a straight line on a log-log scale, i.e., as a power-law function. We believe this is a fair choice since we only look at the tail of the distribution, which looks like a straight line, and since we interpolate between each pair of closest two points only, instead of fitting on the whole tail of the distribution.

The DD of a random network built with our model is presented in Figure 4.3b. For time computation reasons, the built network only has $N = 2 \cdot 10^5$ nodes, to be compared to the $5 \cdot 10^8$ nodes of Twitter. However, it is enough to verify that its DD shape follows the one of the real Twitter's DD: in particular we recognize the spikes around $d = 20$ and $d = 2000$.

4.6 Link between the attachment function and heavy-tailed distributions

In this Section, we show a correlation between the shape of the attachment function f and the tail of the probability function P. More precisely, we show that (under some conditions on f), if f verifies $lim_{i\to+\infty} f(i) = +\infty$, then the associated distribution P is heavy-tailed, and if f is bounded from above, then, the associated distribution P is not heavy-tailed.

The heavy-tailed feature of DDs is an interesting property of networks: most of the time, real-world networks exhibit heavy-tailed DDs, while pure randomness (as we find in the Erdos-Reyni model) build networks with homogeneous DDs. The particular case of linear preferential attachment is known to build networks with heavy-tailed DDs. To the best of our knowledge, this is the first time such a general correlation is made between the attachment function of random growing models and the heavy-tailed feature of the DD. Moreover, if the results presented here only apply for the model proposed in Section 4.3, we believe the proofs can be extended to most of random growing models to show similar results.

Note that we now consider the model in which we impose an attachment function f, and we study the shape of the DD (instead of imposing a probability distribution and study the attachment function, as we have made until now).

4.6.1 Conditions on f

First of all, f has to verify some conditions in order to give a coherent probability distribution. For instance, choosing $f(i) = i^\alpha$ with $\alpha > 1$ build a graph in which a dominant vertex emerges such that after n time steps, the degree of this node is of order n, while the degrees of all other vertices are bounded [21]. Thus the DD associated with this attachment function is not well-defined. We first express the conditions on f. It can be summed up by:

Condition 1. *In order to obtain a distribution P for the DD verifying $\sum_{k\geq 1} P(k) = 1$ and $\sum_{k\geq 1} kP(k) = \mu$, $\mu \in \mathrm{R}_+^*$, the attachment function f has to verify:*

- *If f converges, $\sum_{i=1}^{+\infty} \frac{(1+\frac{1}{c})^{-i+1}}{f(i)}$ is finite, where $c = \max_{i\geq 1}\Big(f(i)\Big)$;*
- *If f diverges, $\sum_{i=1}^{+\infty} \exp\Big(-\sum_{k=1}^{i} \frac{1}{f(k)}\Big)$ is finite.*

Proof. First, we express the condition $\sum_{k\geq 1} kP(k)$ in an interesting form:

Lemma 9.

$$\sum_{k=1}^{+\infty} f(k)P(k) = \sum_{k=1}^{+\infty} kP(k) \tag{4.49}$$

Proof. Using Equation 4.1, we have:

$$\sum_{k=1}^{+\infty} f(k)P(k) = \sum_{k=1}^{+\infty} \sum_{k'=k+1}^{+\infty} P(k') \tag{4.50}$$

$$= \sum_{k=1}^{+\infty} kP(k) \tag{4.51}$$

□

We believe that this surprising equality may lead to insights into the connection between P and f; we keep this exploration for future works.
We are now left with the study of the convergence of $\sum_{k=1}^{+\infty} P(k)$ and $\sum_{k=1}^{+\infty} f(k)P(k)$. Iterating over Equation 4.22 to express P as a function of f gives:

$$P(i) = P(1) \prod_{k=2}^{i} \frac{f(k-1)}{1+f(k)} \tag{4.52}$$

We can rewrite this expression as:

$$P(i) = P(1)\frac{f(1)}{f(i)} \prod_{k=2}^{i} \frac{f(k)}{1+f(k)} \tag{4.53}$$

$$= P(1)\frac{f(1)}{f(i)} \exp\Big(\ln(\prod_{k=2}^{i} \frac{f(k)}{1+f(k)})\Big) \tag{4.54}$$

$$= P(1)\frac{f(1)}{f(i)} \exp\Big(-\sum_{k=2}^{i} ln(1+\frac{1}{f(k)})\Big). \tag{4.55}$$

From now on we distinguish two cases:

1) f converges:
In this case, $\exists c > 0 / \forall i \geq 1, f(i) \leq c$. We have:

$$P(i) \leq P(1)\frac{f(1)}{f(i)} \exp\Big(-\sum_{k=2}^{i} ln(1+\frac{1}{c})\Big) \tag{4.56}$$

$$\leq P(1)f(1)\frac{(1+\frac{1}{c})^{-i+1}}{f(i)} \tag{4.57}$$

$$\implies \sum_{k=1}^{+\infty} P(i)f(i) \leq P(1)f(1)(c+1). \tag{4.58}$$

So, if f converges, $\sum_{k=1}^{+\infty} f(k)P(k)$ always converges, and, by Lemma 9, the mean of P is finite. Equation 4.57 shows that $\sum_{k\geq 1} \frac{(1+\frac{1}{c})^{-k}}{f(k)}$ has to be finite in order to satisfy the condition on $\sum_{k=1}^{+\infty} P(k)$.

2) f diverges:
Then, we can find i_0 such that $\sum_{k=2}^{i} ln(1+\frac{1}{f(k)}) \underset{i\to+\infty}{\sim} \sum_{k=2}^{i_0} ln(1+\frac{1}{f(k)}) + \sum_{k=i_0+1}^{i} \frac{1}{f(k)}$.

We can rewrite Equation (4.55) as:

$$P(i) \sim P(1)\frac{f(1)}{f(i)} \exp\Big(-\sum_{k=2}^{i_0} ln(1+\frac{1}{f(k)}) + \sum_{k=1}^{i_0} \frac{1}{f(k)} - \sum_{k=1}^{i} \frac{1}{f(k)} \Big) \quad (4.59)$$

$$\sim K_{f,i_0} \frac{1}{f(i)} \exp\Big(-\sum_{k=1}^{i} \frac{1}{f(k)} \Big), \quad (4.60)$$

with K_{f,i_0} a constant depending of f and i_0. Thus, by Lemma 9, the mean of P is finite if and only if the following quantity is finite:

$$\sum_{i=1}^{+\infty} \exp\Big(-\sum_{k=1}^{i} \frac{1}{f(k)} \Big).$$

Note that the other condition, i.e., the convergence of $\sum_{i=1}^{+\infty} \frac{1}{f(i)} \exp\Big(-\sum_{k=1}^{i} \frac{1}{f(k)} \Big)$, is included in the first one. Indeed, since f diverges, there exists a constant i_1 such that $\forall i \geq i_1 0, \frac{1}{f(i)} \leq 1$, and the second condition can be bounded by the first one.

□

It is interesting to note that, for $f(i) \propto i^\alpha$, $\alpha = 1$ is the limit case for which Condition 1 holds, as expected from the results of [21].

4.6.2 Link between the limit of f and heavy-tailed DDs

Definition 2. *[23] We say that a distribution P is heavy-tailed if it decays more slowly than an exponential, i.e.:*

$$\forall t > 0, e^{ti} P(X > i) \underset{i \to +\infty}{\to} +\infty.$$

We show the two following theorems:

Theorem 3. *Let f be an attachment function verifying Condition 1 and such that $\lim_{i \to +\infty} f(i) = +\infty$. Then, the associated distribution P is heavy-tailed.*

Theorem 4. *Let f be an attachment function verifying Condition 1 and such that f is bounded from above by $M > 0$. Then, the associated distribution P is not heavy-tailed.*

To prove those theorems, we will use the following lemma:

Lemma 10. *P is heavy-tailed if and only if*

$$\forall t > 0, \exists i_0 > 0 / \lim_{i \to +\infty} h_{t,i_0}(i) = +\infty,$$

where $h_{t,i_0}(i) = ti + log(f(i_0)) - \sum_{k=i_0}^{i-1} log(1 + \frac{1}{f(k+1)})$.

Proof. We recall that $P(i) = P(1)\prod_{k=1}^{i-1}\frac{f(k)}{1+f(k+1)}$ and $f(i) = \frac{1}{P(i)}\sum_{k=i+1}^{\infty} P(k)$. It implies

$$P(X > i) := \sum_{k=i+1}^{\infty} P(k) = f(i)P(i) = f(i)P(1)\prod_{k=1}^{i-1}\frac{f(k)}{1+f(k+1)}. \tag{4.61}$$

Let $t > 0$, $i_0 > 0$. We have:

$$e^{ti}P(X > i) = e^{ti}f(i)P(1)\prod_{k=1}^{i-1}\frac{f(k)}{1+f(k+1)} \tag{4.62}$$

$$= e^{ti}e^{log(f(i))}P(1)\prod_{k=1}^{i_0-1}\frac{f(k)}{1+f(k+1)}\prod_{k=i_0}^{i-1}e^{log(\frac{f(k)}{1+f(k+1)})} \tag{4.63}$$

$$= P(1)\prod_{k=1}^{i_0-1}\big(\frac{f(k)}{1+f(k+1)}\big) \times e^{ti+log(f(i))+\sum_{k=i_0}^{i-1} log\left(\frac{f(k)}{1+f(k+1)}\right)}. \tag{4.64}$$

We call $h_{t,i_0}(i) = ti + log(f(i)) + \sum_{k=i_0}^{i-1} log(\frac{f(k)}{1+f(k+1)})$. P is heavy-tailed if and only if $\lim_{i\to+\infty} h_{t,i_0}(i) = +\infty$. But h_{t,i_0} can also be expressed as:

$$h_{t,i_0}(i) = ti + \log(f(i)) - \sum_{k=i_0}^{i-1}\log\left(\frac{1+f(k+1)}{f(k)}\right) \tag{4.65}$$

$$= ti + \log(f(i)) - \sum_{k=i_0}^{i-1}\log\left(\frac{f(k+1)}{f(k)}(1+\frac{1}{f(k+1)})\right) \tag{4.66}$$

$$= ti + \log(f(i)) - \sum_{k=i_0}^{i-1}\log(f(k+1)) + \sum_{k=i_0}^{i-1}\log(f(k)) - \sum_{k=i_0}^{i-1}\log\left(1+\frac{1}{f(k+1)}\right) \tag{4.67}$$

$$= ti + \log(f(i)) - \log(f(i)) + \log(f(i_0)) - \sum_{k=i_0}^{i-1}\log\left(1+\frac{1}{f(k+1)}\right) \tag{4.68}$$

$$= ti + log(f(i_0)) - \sum_{k=i_0}^{i-1}\log\left(1+\frac{1}{f(k+1)}\right). \tag{4.69}$$

□

Proof of Theorem 3.

Let $t > 0$. By definition of the limit, $\exists i_0 / \forall i > i_0, f(i) > \frac{1}{e^{t/2}-1}$. So:

$$h_{t,i_0}(i) = ti + log(f(i_0)) - \sum_{k=i_0}^{i-1} \log\left(1 + \frac{1}{f(k+1)}\right) \tag{4.70}$$

$$> ti + \log(f(i_0)) - \sum_{k=i_0}^{i-1} \log\left(1 + \frac{1}{(\frac{1}{e^{t/2}-1})}\right) \tag{4.71}$$

$$= ti + \log(f(i_0)) - (i - i_0 - 1)\frac{t}{2} \tag{4.72}$$

$$= \frac{t}{2}i + log(f(i_0)) + (i_0 + 1)\frac{t}{2} \tag{4.73}$$

$$\underset{i\to+\infty}{\to} +\infty. \tag{4.74}$$

□

Proof of Theorem 4.

$$h_{t,i_0}(i) = ti + \log(f(i_0)) - \sum_{k=i_0}^{i-1} log\left(1 + \frac{1}{f(k+1)}\right) \tag{4.75}$$

$$< ti + \log(f(i_0)) - \sum_{k=i_0}^{i-1} \log\left(1 + \frac{1}{M}\right) \tag{4.76}$$

$$= ti + \log(f(i_0)) - (i - i_0 - 1)\log\left(1 + \frac{1}{M}\right). \tag{4.77}$$

Let $t = \frac{1}{2}log(1 + \frac{1}{M})$.

$$h_{t,i_0}(i) = -\frac{1}{2}\log\left(1 + \frac{1}{M}\right) i + \log(f(i_0)) + (i_0 + 1)\log(1 + \frac{1}{M}) \tag{4.78}$$

$$\underset{i\to+\infty}{\to} -\infty. \tag{4.79}$$

There exists a value of $t > 0$ such that the limit of h_{t,i_0} goes to $-\infty$ for any i_0, hence P is not heavy-tailed. □

Remark 4. *The set of* preferential attachment functions *(i.e., increasing functions) is not included nor it contains any of previous cases. Indeed, we can have a preferential attachment (or a non preferential attachment function) in the first case, as well as in the second case. Indeed, bounded increasing functions are covered by Theorem 4 while non bounded increasing functions are covered by Theorem 3.*

Remark 5. *Not all functions are included in the previous cases. It remains the cases where the limit of f is not infinite but f is not bounded either (for instance, $f(i) = 1$ if i is pair, $f(i) = i$ otherwise). However, we believe those cases are hardly encountered in practice.*

4.7 Conclusion

In this Chapter, we proposed a new random growth model picking the nodes to be connected together in the graph with a flexible probability f. We expressed this

f as a function of any distribution P, leading to the possibility to build a random network with any wanted degree distribution. We computed f for some classical distributions, as much as for a snapshot of Twitter of 505 million nodes and 23 billion edges. We believe this model is useful for anyone studying networks with atypical degree distributions, regardless of the domain. If the presented model is undirected, we also believe a directed version of it, based on the Bollobás et al. model [4], can be easily generalized from the presented one. We also believe this model can enlighten relations between the degree distributions of networks and the attachment function behind them, both in random growth models as well as real-world networks. To take a step in that direction, we show that, in our model, the limit of the attachment function f is sufficient to determine if the probability distribution of the graphs is heavy-tailed or not. We believe this result can be extended to other models, and hopefully lead to interesting studies on real-world networks.

Bibliography

[1] Yong-Yeol Ahn, Seungyeop Han, Haewoon Kwak, Sue Moon, and Hawoong Jeong. Analysis of topological characteristics of huge online social networking services. In *Proceedings of the 16th int. conference on World Wide Web*, pages 835–844, 2007.

[2] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47, 2002.

[3] Béla Bollobás. A probabilistic proof of an asymptotic formula for the number of labelled regular graphs. *European Journal of Combinatorics*, 1(4):311–316, 1980.

[4] Béla Bollobás, Christian Borgs, Jennifer T Chayes, and Oliver Riordan. Directed scale-free graphs. In *SODA*, volume 3, pages 132–139, 2003.

[5] Anna D Broido and Aaron Clauset. Scale-free networks are rare. *Nature communications*, 10(1):1–10, 2019.

[6] Meeyoung Cha, Alan Mislove, and Krishna P Gummadi. A measurement-driven analysis of information propagation in the flickr social network. In *Proceedings of the 18th international conference on World wide web*, pages 721–730, 2009.

[7] Fan Chung, Fan RK Chung, Fan Chung Graham, Linyuan Lu, Kian Fan Chung, et al. *Complex graphs and networks*. American Mathematical Soc., 2006.

[8] Fan Chung and Linyuan Lu. *Complex Graphs and Networks*. American Mathematical Society, 2006.

[9] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

[10] Paul Erdős and Alfréd Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5(1):17–60, 1960.

[11] Maksym Gabielkov and Arnaud Legout. The complete picture of the twitter social graph. In *Proc. on CoNEXT student workshop*, pages 19–20. ACM, 2012.

[12] Gourab Ghoshal and MEJ Newman. Growing distributed networks with arbitrary degree distributions. *The European Physical Journal B*, 58(2):175–184, 2007.

[13] Minas Gjoka, Maciej Kurant, Carter T Butts, and Athina Markopoulou. Walking in facebook: A case study of unbiased sampling of osns. In *IEEE INFOCOM*, 2010.

[14] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301), 1963.

[15] Gudlaugur Jóhannesson, Gunnlaugur Björnsson, and Einar H Gudmundsson. Afterglow light curves and broken power laws: a statistical study. *The Astrophysical Journal Letters*, 640(1):L5, 2006.

[16] Jure Leskovec and Eric Horvitz. Planetary-scale views on a large instant-messaging network. In *Proc. of the 17th international conference on World Wide Web*, 2008.

[17] Gipsi Lima-Mendez and Jacques van Helden. The powerful law of the power law and other myths in network biology. *Molecular BioSystems*, 5(12):1482–1493, 2009.

[18] Michael Mitzenmacher and Eli Upfal. *Probability and Computing: Randomization and Probabilistic Techniques in Algorithms and Data Analysis*. Cambridge University Press, USA, 2nd edition, 2017.

[19] Seth A Myers, Aneesh Sharma, Pankaj Gupta, and Jimmy Lin. Information network or social network?: the structure of the twitter follow graph. In *Proceedings of the 23rd Int. Conference on World Wide Web*, pages 493–498. ACM, 2014.

[20] Mark EJ Newman, Steven H Strogatz, and Duncan J Watts. Random graphs with arbitrary degree distributions and their applications. *Physical review E*, 2001.

[21] Roberto Oliveira and Joel Spencer. Connectivity transitions in networks with super-linear preferential attachment. *Internet Mathematics*, 2(2):121–163, 2005.

[22] Nataša Pržulj. Biological network comparison using graphlet degree distribution. *Bioinformatics*, 23(2):e177–e183, 2007.

[23] Tomasz Rolski, Hanspeter Schmidli, Volker Schmidt, and Jozef L Teugels. *Stochastic processes for insurance and finance*, volume 505. John Wiley & Sons, 2009.

[24] Arnaud Sallaberry, Faraz Zaidi, and Guy Melançon. Model for generating artificial social networks having community structures with small-world and scale-free properties. *Social Network Analysis and Mining*, 3(3):597–609, 2013.

[25] Mukund Seshadri, Sridhar Machiraju, Ashwin Sridharan, Jean Bolot, Christos Faloutsos, and Jure Leskove. Mobile call graphs: beyond power-law and lognormal distributions. In *ACM SIGKDD*, pages 596–604, 2008.

[26] Andrew T Stephen and Olivier Toubia. Explaining the power-law degree distribution in a social commerce network. *Social Networks*, 31(4):262–270, 2009.

[27] Thibaud Trolliet, Nathann Cohen, Frédéric Giroire, Luc Hogie, and Stéphane Pérennes. Interest clustering coefficient: a new metric for directed networks like twitter. *arXiv preprint arXiv:2008.00517*, 2020.

Chapter 5

Preferential attachment hypergraph with high modularity

5.1 Introduction

We recall that it was empirically recognized that the common ground of large complex networks are small diameter, high clustering coefficient, heavy tailed degree distribution and visible community structure [5]. Surprisingly, all those characteristics appear, no matter whether we investigate biological, social, or technological systems.

A number of theoretical models were presented throughout last 25 years. Just to mention the mostly investigated ones: Watts and Strogatz (exhibiting small-world and high clustering properties [26]), Molloy and Reed (with a given degree sequence [20]), Chung-Lu (with a given expected degree sequence [8]), Cooper-Frieze (model of web graphs [10]), Buckley-Osthus [7] or random intersection graph (with high clustering properties and following a power-law, [4]). None of here mentioned graphs captures all the four properties listed in the previous paragraph, e.g., [26] does not have a heavy tailed degree distribution, [2] and [8] models suffer from vanishing clustering coefficient [5], almost all of them do not exhibit visible community structure, i.e., have low *modularity*.

Modularity is a parameter measuring how clearly a network may be divided into *communities*. It was introduced by Newman and Girvan in [22]. A graph has high modularity if it is possible to partition the set of its vertices into communities inside which the density of edges is remarkably higher than the density of edges between different communities. Modularity is known to have some drawbacks (for thorough discussion check [18]). Nevertheless, today it remains a popular measure and is widely used in most common algorithms for community detection [12, 3, 24]. It is well known that the real-life social or biological networks are highly modular [11, 13]. At the same time simulations show that most of existing preferential attachment models have low modularity. Good modularity properties one finds in geometric models, like spatial preferential attachment graphs [16, 15], however they use additionally a spatial metric.

Finally, almost all the up-to-date complex networks models are graph models thus are able to mirror only binary relations. In practical applications k-ary relations (co-authorship, groups of interests or protein reactions) are often modelled in graphs by cliques which may lead to a profound information loss.

Results. Within this article we propose a dynamic model with high modularity by preserving a heavy tailed degree distribution and not using a spatial metric. Moreover, our model is a random hypergraph (not a graph) thus can reflect k-ary relations. Preferential attachment hypergraph model was first introduced by Wang et al. in [25]. However, it was restricted just to a specific subfamily of uniform acyclic hypergraphs (the analogue of trees within graphs). The first rigorously studied non-uniform hypergraph preferential attachment model was proposed only in 2019 by Avin et al. [1]. Its degree distribution follows a power-law. However, our empirical results indicate that this model has a weakness of low modularity (see Section 5.7.2). To the best of our knowledge the model proposed within this article is the first dynamic non-uniform hypergraph model with degree sequence following a power-law and exhibiting clear community structure. We experimentally show that features of our model correspond to the ones of a real co-authorship network built upon Scopus database.

Chapter organisation. Basic definitions are introduced in Sec. 5.2. In Sec. 5.3, we present a universal preferential attachment hypergraph model which unifies many existing models (from classical Barabási-Albert graph [2] to Avin et al. preferential attachment hypergraph [1]). In Sec. 5.4, we use it as a component in a stochastic block model to build a general hypergraph with good modularity properties. Theoretical bounds for its modularity and experimental results on a real data are presented in Sec. 5.7. Further works are presented in Sec. 5.8.

5.2 Basic definitions and notation

We define a *hypergraph* H as a pair $H = (V, E)$, where V is a set of vertices and E is a set of hyperedges, i.e., non-empty, unordered multisets of V. We allow for a multiple appearance of a vertex in a hyperedge (self-loops). The degree of a vertex v in a hyperedge e, denoted by $d(v, e)$, is the number of times v appears in e. The cardinality of a hyperedge e is $|e| = \sum_{v \in e} d(v, e)$. The degree of a vertex $v \in V$ in H is understood as the number of times it appears in all hyperedges, i.e., $\deg(v) = \sum_{e \in E} d(v, e)$. If $|e| = k$ for all $e \in E$, H is said *k-uniform*.

We consider hypergraphs that grow by adding vertices and/or hyperedges at discrete time steps $t = 0, 1, 2, \ldots$. The hypergraph obtained at time t will be denoted by $H_t = (V_t, E_t)$ and the degree of $u \in V_t$ in H_t by $\deg_t(u)$. By D_t we denote the sum of degrees at time t, i.e., $D_t = \sum_{u \in V_t} \deg_t(u)$. As the hypergraph gets large, the probability of creating a self-loop can be well bounded and is quite small provided that the sizes of hyperedges are reasonably bounded.

$N_{k,t}$ stands for the number of vertices in H_t of degree k. We say that the degree distribution of a hypergraph follows a *power-law* if the fraction of vertices of degree k is proportional to $k^{-\beta}$ for some exponent $\beta \geq 1$. Formally, we will interpret it as $\lim_{t\to\infty} \mathbb{E}\left[\frac{N_{k,t}}{|V_t|}\right] \sim c \cdot k^{-\beta}$ for some positive constant c and $\beta \geq 1$. For f and g being real functions we write $f(k) \sim g(k)$ if $f(k)/g(k) \xrightarrow{k\to\infty} 1$.

Modularity measures the presence of community structure in the graph. Its definition for graphs introduced by Newman and Girvan in 2004 is given below.

Definition 3 ([22])**.** *Let $G = (V, E)$ be a graph with at least one edge. For a*

partition $\mathcal{A}$ of vertices of G define its modularity score on G as

$$q_{\mathcal{A}}(G) = \sum_{A \in \mathcal{A}} \left(\frac{|E(A)|}{|E|} - \left(\frac{vol(A)}{2|E|} \right)^2 \right),$$

where $E(A)$ is the set of edges within A and $vol(A) = \sum_{v \in A} deg(v)$. Modularity of G is given by $q^*(G) = \max_{\mathcal{A}} q_{\mathcal{A}}(G)$.

Conventionally, a graph with no edges has modularity equal to 0. The value $\sum_{A \in \mathcal{A}} \frac{|E(A)|}{|E|}$ is called an *edge contribution* while $\sum_{A \in \mathcal{A}} \left(\frac{vol(A)}{2|E|} \right)^2$ is a *degree tax*. A single summand of the modularity score is the difference between the fraction of edges within A and the expected fraction of edges within A if we considered a random multigraph on V with the degree sequence given by G. One can observe that the value of $q^*(G)$ always falls into the interval $[0, 1)$.

Several approaches to define a modularity for hypergraphs can be found in contemporary literature. Some of them flatten a hypergraph to a graph (e.g., by replacing each hyperedge by a clique) and apply a modularity for graphs (see e.g. [21]). Others base on information entropy modularity [27]. We want to stick to the classical definition from [22] and preserve a rich hypergraph structure, therefore we work with the definition proposed by Kamiński et al. in [17].

Definition 4 ([17])**.** *Let $H = (V, E)$ be a hypergraph with at least one hyperedge. For $\ell \geq 1$ let $E_\ell \subseteq E$ denote the set of hyperedges of cardinality ℓ. For a partition $\mathcal{A}$ of vertices of H define its modularity score on H as*

$$q_{\mathcal{A}}(H) = \sum_{A \in \mathcal{A}} \left(\frac{|E(A)|}{|E|} - \sum_{\ell \geq 1} \frac{|E_\ell|}{|E|} \cdot \left(\frac{vol(A)}{vol(V)} \right)^\ell \right),$$

where $E(A)$ is the set of hyperedges within A (a hyperedge is within A if all its vertices are contained in A), $vol(A) = \sum_{v \in A} deg(v)$ and $vol(V) = \sum_{v \in V} deg(v)$. Modularity of H is given by $q^*(H) = \max_{\mathcal{A}} q_{\mathcal{A}}(H)$.

A single summand of the degree tax is the expected number of hyperedges within A if we considered a random hypergraph on V with the degree sequence given by H and having the same number of hyperedges of corresponding cardinalities.

We write that an event A occurs *with high probability* (whp) if the probability $\mathbb{P}[A]$ depends on a certain number t and tends to 1 as t tends to infinity.

5.3 General preferential attachment hypergraph model

In this section we generalise a hypergraph model proposed by Avin et al. in [1]. Model from [1] allows for two different actions at a single time step - attaching a new vertex by a hyperedge to the existing structure or creating a new hyperedge on already existing vertices. We allow for four different events at a single time step, admit the possibility of adding more than one hyperedge at once and draw the cardinality of newly created hyperedge from more than one distribution. The events allowed at a single time step in our model H_t are: adding an isolated vertex,

adding a vertex and attaching it to the existing structure by m hyperedges, adding m hyperedges, or doing nothing. The last event "doing nothing" is included since later we put H_t in a broader context of stochastic block model, where it serves as a single community. "Doing nothing" indicates a time slot in which nothing associated directly with H_t happens but some event takes place in the other part of the whole stochastic block model.

5.3.1 Model $\mathbf{H(H_0, p, Y, X, m, \gamma)}$

General hypergraph model H is characterized by six parameters. These are:

1. H_0 - initial hypergraph, seen at $t = 0$;

2. $\mathbf{p} = (p_v, p_{ve}, p_e)$ - vector of probabilities indicating, what are the chances that a particular type of event occurs at a single time step; we assume $p_v + p_{ve} + p_e \in (0, 1]$; additionally p_e is split into the sum of r probabilities $p_e = p_e^{(1)} + p_e^{(2)} + \ldots + p_e^{(r)}$ which allows for adding hyperedges whose cardinalities follow different distributions;

3. $Y = (Y_0, Y_1, \ldots, Y_t, \ldots)$ - independent random variables, cardinalities of hyperedges that are added together with a vertex at a single time step;

4. $X = ((X_1^{(1)}, \ldots, X_t^{(1)}, \ldots), (X_1^{(2)}, \ldots, X_t^{(2)}, \ldots), \ldots, (X_1^{(r)}, \ldots, X_t^{(r)}, \ldots))$ - r sequences of independent random variables, cardinalities of hyperedges that are added at a single time step when no new vertex is added;

5. m - number of hyperedges added at once;

6. $\gamma \geqslant 0$ - parameter appearing in the formula for the probability of choosing a particular vertex to a newly created hyperedge.

Here is how the structure of $H = H(H_0, p, Y, X, m, \gamma)$ is being built. We start with some non-empty hypergraph H_0 at $t = 0$. We assume for simplicity that H_0 consists of a hyperedge of cardinality 1 over a single vertex. Nevertheless, all the proofs may be generalised to any initial H_0 having constant number of vertices and constant number of hyperedges with constant cardinalities. 'Vertices chosen from V_t in proportion to degrees' means that vertices are chosen independently (possibly with repetitions) and the probability that any u from V_t is chosen is

$$\mathbb{P}[u \text{ is chosen}] = \frac{\deg_t(u) + \gamma}{\sum_{v \in V_t}(\deg_t(v) + \gamma)} = \frac{\deg_t(u) + \gamma}{D_t + \gamma|V_t|}.$$

For $t \geqslant 0$ we form H_{t+1} from H_t choosing only one of the following events according to $\mathbf{p}$.

- With probability p_v: Add one new isolated vertex.

- With probability p_{ve}: Add one vertex v. Draw a value y being a realization of Y_t. Then repeat m times: select $y - 1$ vertices from V_t in proportion to degrees; add a new hyperedge consisting of v and $y - 1$ selected vertices.

- With probability $p_e^{(1)}$: Draw a value x being a realization of $X_t^{(1)}$. Then repeat m times: select x vertices from V_t in proportion to degrees; add a new hyperedge consisting of x selected vertices.
- ...
- With probability $p_e^{(r)}$: Draw a value x being a realization of $X_t^{(r)}$. Then repeat m times: select x vertices from V_t in proportion to degrees; add a new hyperedge consisting of x selected vertices.
- With probability $1 - (p_v + p_{ve} + p_e)$: Do nothing.

We allow for r different distributions from which one can draw the cardinality of newly created hyperedges. Later, when H_t serves as a single community in the context of the whole stochastic block model, this trick allows for spanning a new hyperedge across several communities drawing vertices from each of them according to different distributions. This reflects some possible real-life applications. Think of an article authored by people from two different research centers. Our experimental observation is that it is very unlikely that the number of authors will be distributed uniformly among two centers. More often, one author represents one center, while the others are affiliated with the second one.

5.3.2 Degree distribution of $\mathbf{H(H_0, p, Y, X, m, \gamma)}$

In this section we prove that the degree distribution of $H = H(H_0, p, Y, X, m, \gamma)$ follows a power-law with $\beta > 2$. We assume that supports of random variables indicating cardinalities of hyperedges are bounded and their expectations are constant. This assumption is in accord with potential applications - think of co-authors, groups of interest, protein reactions, ect.

Theorem 5. *Consider a hypergraph $H = H(H_0, \mathbf{p}, Y, X, m, \gamma)$ for any $t > 0$. Let $i \in \{1, \ldots, r\}$. Let $\mathbb{E}[Y_t] = \mu_0$, and $\mathbb{E}[X_t^{(i)}] = \mu_i$. Moreover, let $1 \leqslant Y_t < t^{1/4}$ and $1 \leqslant X_t^{(i)} < t^{1/4}$. Then the degree distribution of H follows a power-law with*

$$\beta = 2 + \frac{\gamma \bar{V} + m \cdot p_{ve}}{\bar{D} - m \cdot p_{ve}},$$

where $\bar{V} = p_v + p_{ve}$ and $\bar{D} = m(p_{ve}\mu_0 + p_e^{(1)}\mu_1 + \ldots + p_e^{(r)}\mu_r)$ which are the expected number of vertices added per a single time step and the expected number of vertices that increase their degree in a single time step, respectively.

The number of vertices in H_t is a random variable following a binomial distribution. Since $|V_0| = 1$ we have $|V_t| \sim B(t, p_v + p_{ve}) + 1$. Since $|E_0| = 1$, the number of hyperedges in H_t is a random variable satisfying $|E_t| \sim mB(t, p_{ve} + p_e) + 1$.

Before we prove Theorem 5 we discuss briefly the concentration of random variables $|V_t|$ (the number of vertices at time t), D_t (the sum of degrees at time t) and $W_t = D_t + \gamma|V_t|$. We also state two technical lemmas that will be helpful later on.

Lemma 11 (Chernoff bounds, [19], Chapter 4.2)**.** *Let $Z_1, Z_2, \ldots, Z_t$ be independent indicator random variables with $\mathbb{P}[Z_i = 1] = p_i$ and $\mathbb{P}[Z_i = 0] = 1 - p_i$. Let $\delta > 0$, $Z = \sum_{i=1}^t Z_i$ and $\mu = \mathbb{E}[Z] = \sum_{i=1}^t p_i$. Then*

$$\mathbb{P}[|Z - \mu| \geqslant \delta\mu] \leq 2e^{-\mu\delta^2/3}.$$

Corollary 2. *Since* $|V_t| \sim B(t, p_v + p_{ve}) + 1$ *setting* $\delta = \sqrt{\frac{9 \ln t}{(p_v + p_{ve})t}}$ *in Lemma 11 we get*

$$\mathbb{P}[||V_t| - \mathbb{E}[]|V_t|]| \geqslant \sqrt{9(p_v + p_{ve})t \ln t}] \leq 2/t^3.$$

Lemma 12. *If* $\lim_{t\to\infty} \frac{\mathbb{E}[N_{k,t}]}{t} \sim ck^{-\beta}$ *for some positive constant* c *then*

$$\lim_{t\to\infty} \mathbb{E}\left[\frac{N_{k,t}}{|V_t|}\right] \sim \frac{c}{p_v + p_{ve}} k^{-\beta}.$$

(*Here "*$\sim$*" refers to the limit by* $k \to \infty$.)

Proof. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be the probability space on which random variables $N_{k,t}$ and $|V_t|$ are defined. Thus $N_{k,t} : \Omega \to \mathbb{R}$ and $|V_t| : \Omega \to \mathbb{R}$. Let $\Omega_1 \subseteq \Omega$ denote the set of all $\omega \in \Omega$ such that $|V_t|(\omega) \in (\mathbb{E}|V_t| - \sqrt{9(p_v + p_{ve})t \ln t}, \mathbb{E}|V_t| + \sqrt{9(p_v + p_{ve})t \ln t})$. By Corollary 2 we know that $\sum_{\omega\in\Omega\setminus\Omega_1} \mathbb{P}[\omega] \leqslant 2/t^3$. Using the fact that for each ω $\frac{N_{k,t}(\omega)}{|V_t|(\omega)} \leqslant 1$ we get

$$\mathbb{E}\left[\frac{N_{k,t}}{|V_t|}\right] = \sum_{\omega\in\Omega} \frac{N_{k,t}(\omega)}{|V_t|(\omega)} \mathbb{P}[\omega] = \sum_{\omega\in\Omega_1} \frac{N_{k,t}(\omega)}{|V_t|(\omega)} \mathbb{P}[\omega] + \sum_{\omega\in\Omega\setminus\Omega_1} \frac{N_{k,t}(\omega)}{|V_t|(\omega)} \mathbb{P}[\omega] \tag{5.1}$$

$$\leqslant \sum_{\omega\in\Omega} \frac{N_{k,t}(\omega)}{\mathbb{E}|V_t| - \sqrt{9(p_v + p_{ve})t \ln t}} \mathbb{P}[\omega] + \sum_{\omega\in\Omega\setminus\Omega_1} 1 \cdot \mathbb{P}[\omega] \tag{5.2}$$

$$\leqslant \frac{\mathbb{E}[N_{k,t}]}{\mathbb{E}|V_t| - \sqrt{9(p_v + p_{ve})t \ln t}} + 2/t^3 \sim \frac{\mathbb{E}[N_{k,t}]}{(p_v + p_{ve})t}. \tag{5.3}$$

On the other hand, since $N_{k,t} \leq t$,

$$\mathbb{E}\left[\frac{N_{k,t}}{|V_t|}\right] \geqslant \sum_{\omega\in\Omega_1} \frac{N_{k,t}(\omega)}{|V_t|(\omega)} \mathbb{P}[\omega] \geqslant \sum_{\omega\in\Omega_1} \frac{N_{k,t}(\omega)}{\mathbb{E}|V_t| + \sqrt{9(p_v + p_{ve})t \ln t}} \mathbb{P}[\omega] \tag{5.4}$$

$$= \frac{1}{\mathbb{E}|V_t| + \sqrt{9(p_v + p_{ve})t \ln t}} \left(\mathbb{E}[N_{k,t}] - \sum_{\omega\in\Omega\setminus\Omega_1} N_{k,t}(\omega) \mathbb{P}[\omega]\right) \tag{5.5}$$

$$\geqslant \frac{1}{\mathbb{E}|V_t| + \sqrt{9(p_v + p_{ve})t \ln t}} \left(\mathbb{E}[N_{k,t}] - \sum_{\omega\in\Omega\setminus\Omega_1} t \cdot \mathbb{P}[\omega]\right) \tag{5.6}$$

$$\geqslant \frac{\mathbb{E}[N_{k,t}]}{\mathbb{E}|V_t| + \sqrt{9(p_v + p_{ve})t \ln t}} - \frac{t \cdot 2/t^3}{\mathbb{E}|V_t| + \sqrt{9(p_v + p_{ve})t \ln t}} \tag{5.7}$$

$$\sim \frac{\mathbb{E}[N_{k,t}]}{(p_v + p_{ve})t}. \tag{5.8}$$

□

Lemma 13 (Hoeffding's inequality, [14])**.** *Let* $Z_1, Z_2, \ldots, Z_t$ *be independent random variables such that* $\mathbb{P}[Z_i \in [a_i, b_i]] = 1$. *Let* $\delta > 0$ *and* $Z = \sum_{i=1}^t Z_i$. *Then*

$$\mathbb{P}[|Z - \mathbb{E}[Z]| \geqslant \delta] \leqslant 2 \exp\left\{-\frac{2\delta^2}{\sum_{i=1}^t (a_i - b_i)^2}\right\}.$$

Lemma 14. *Let $t > 0$. Let $\mathbb{E}[Y_t] = \mu_0$, and $\mathbb{E}[X_t^{(i)}] = \mu_i$ for $i \in \{1, 2, \ldots, r\}$. Moreover, let $2 \leqslant Y_t < t^{1/4}$ and $1 \leqslant X_t^{(i)} < t^{1/4}$ for $i \in \{1, 2, \ldots, r\}$. Let $W_t = D_t + \gamma|V_t|$. Then*

$$\mathbb{P}[|W_t - \mathbb{E}[W_t]| \geq mt^{3/4}\sqrt{2\ln t}] = \mathcal{O}\left(\frac{1}{t^4}\right).$$

Proof. Our initial hypergraph consists of a single hyperedge of cardinality 1 over a single vertex thus $W_0 = \gamma + 1$. For $t \geq 1$ we can obtain W_t from W_{t-1} by adding:

1. either γ with probability p_v,
2. or $\gamma + mY_t$ with probability p_{ve},
3. or $mX_t^{(1)}$ with probability $p_e^{(1)}$,
4. or $mX_t^{(2)}$ with probability $p_e^{(2)}$,
5. or ...,
6. or $mX_t^{(r)}$ with probability $p_e^{(r)}$,
7. or 0 with probability $1 - p_v - p_{ve} - p_e$.

Thus we can express W_t as the sum of independent random variables $W_t = \gamma + 1 + Z_1 + Z_2 + \ldots + Z_t$, where $\mathbb{E}[Z_i] = \gamma\bar{V} + \bar{D}$ and $1 \leqslant Z_i \leqslant mt^{1/4} + \gamma$ for $i \in \{1, 2, \ldots, t\}$ and $\bar{D}$ and $\bar{V}$ are defined as in Theorem 5:

$$\bar{V} = p_v + p_{ve} \quad \text{and} \quad \bar{D} = m(p_{ve}\mu_0 + p_e^{(1)}\mu_1 + \ldots + p_e^{(r)}\mu_r).$$

Now, setting $\delta = mt^{3/4}\sqrt{2\ln t}$ in Hoeffding's inequality (see Lemma 13) we get

$$\mathbb{P}[|W_t - \mathbb{E}[W_t]| \geqslant mt^{3/4}\sqrt{2\ln t}] \leqslant 2\exp\left\{-\frac{4 \cdot m^2 \cdot t^{6/4} \cdot \ln t}{(t+1)(m \cdot t^{1/4} + \gamma)^2}\right\} = \mathcal{O}\left(\frac{1}{t^4}\right).$$

□

Lemma 15 ([9], Chapter 3.3)**.** *Let $\{a_t\}$ be a sequence satisfying the recursive relation*

$$a_{t+1} = \left(1 - \frac{b_t}{t}\right)a_t + c_t$$

where $b_t \xrightarrow{t\to\infty} b > 0$ and $c_t \xrightarrow{t\to\infty} c$. Then the limit $\lim_{t\to\infty} \frac{a_t}{t}$ exists and

$$\lim_{t\to\infty} \frac{a_t}{t} = \frac{c}{1+b}.$$

Now we are ready to prove Theorem 5.

Theorem 5. Here we take a standard master equation approach that can be found e.g. in Chung and Lu book [9] about complex networks or Avin et al. paper [1] on preferential attachment hypergraphs.

Recall that $N_{k,t}$ denotes the number of vertices of degree k at time t. We need to show that

$$\lim_{t\to\infty} \mathbb{E}\left[\frac{N_{k,t}}{|V_t|}\right] \sim \tilde{c}k^{-\beta} \tag{5.9}$$

for some constant $\tilde{c}$ and $\beta = 2 + \frac{\gamma\bar{V}+m\cdot p_{ve}}{\bar{D}-m\cdot p_{ve}}$. However, by Lemma 12 we know that it suffices to show that

$$\lim_{t\to\infty} \frac{\mathbb{E}[N_{k,t}]}{t} \sim ck^{-\beta} \tag{5.10}$$

for some constant c.

Our initial hypergraph H_0 consists of a single hyperedge of cardinality 1 over a single vertex thus we can write $N_{0,0} = 0$ and $N_{1,0} = 1$. Now, to formulate a recurrent master equation we make the following observation for $t \geqslant 1$. The vertex v has degree k at time t if either it had degree k at time $t-1$ and was not chosen to any new hyperedge or it had degree $k-i$ at time $t-1$ and was chosen i times to new hyperedges. Note that i can be at most $\min\{k, mZ_t\}$, where Z_t represents a random variable chosen among $Y_t, X_t^{(1)}, \ldots, X_t^{(r)}$ according to $(p_v, p_{ve}, p_e^{(1)}, \ldots, p_e^{(r)})$. Additionally, at each time step a vertex of degree 0 may appear as the new one with probability p_v and a vertex of degree m may appear as the new one with probability p_{ve}. Let $\mathcal{F}_t$ be the σ-algebra associated with the probability space at time t. Let $Q_{d,k,t}$ denote the probability that a specific vertex of degree k was chosen d times to be included in new hyperedges at time t (this probability is expressed as a random variable since it depends on a specific realization of the process up to time $t-1$). Let also $W_t = D_t + \gamma|V_t|$. For $t \geqslant 1$ we get

$$\mathbb{E}[N_{0,t}|\mathcal{F}_{t-1}] = p_v + N_{0,t-1}Q_{0,0,t} \tag{5.11}$$

and when $k \geqslant 1$

$$\begin{aligned}\mathbb{E}[N_{k,t}|\mathcal{F}_{t-1}] = \delta_{k,m}p_{ve} + N_{k,t-1}Q_{0,k,t} + N_{k-1,t-1}Q_{1,k-1,t} \\ + \sum_{i=2}^{\min\{k,mZ_t\}} N_{k-i,t-1}Q_{i,k-i,t},\end{aligned} \tag{5.12}$$

where $\delta_{k,m}$ is the Kronecker delta. We have extracted the first two terms in the above sum since below we prove that these are the dominating terms. Taking expectation on both sides we obtain

$$\mathbb{E}[N_{0,t}] = p_v + \mathbb{E}[N_{0,t-1}Q_{0,0,t}] \tag{5.13}$$

and for $k \geqslant 1$

$$\begin{aligned}\mathbb{E}[N_{k,t}] = \delta_{k,m}p_{ve} + \mathbb{E}[N_{k,t-1}Q_{0,k,t}] + \mathbb{E}[N_{k-1,t-1}Q_{1,k-1,t}] \\ + \sum_{i=2}^{\min\{k,mZ_t\}} \mathbb{E}[N_{k-i,t-1}Q_{i,k-i,t}].\end{aligned} \tag{5.14}$$

Note that

$$\begin{aligned}Q_{0,k,t} = p_v + (1 - p_v - p_{ve} - p_e) + p_{ve}\mathbb{E}\left[\left(1 - \frac{k+\gamma}{W_{t-1}}\right)^{m(Y_t-1)} |\mathcal{F}_{t-1}\right] \\ + p_e^{(1)}\mathbb{E}\left[\left(1 - \frac{k+\gamma}{W_{t-1}}\right)^{mX_t^{(1)}} |\mathcal{F}_{t-1}\right] + \ldots \\ + p_e^{(r)}\mathbb{E}\left[\left(1 - \frac{k+\gamma}{W_{t-1}}\right)^{mX_t^{(r)}} |\mathcal{F}_{t-1}\right]\end{aligned} \tag{5.15}$$

while for $i \in \{1, 2, \ldots, k\}$

$$\begin{aligned} Q_{i,k-i,t} = p_{ve}\mathbb{E}\left[\binom{m(Y_t-1)}{i}\left(\frac{k-i+\gamma}{W_{t-1}}\right)^i\left(1-\frac{k-i+\gamma}{W_{t-1}}\right)^{m(Y_t-1)-i}|\mathcal{F}_{t-1}\right] \\ + p_e^{(1)}\mathbb{E}\left[\binom{mX_t^{(1)}}{i}\left(\frac{k-i+\gamma}{W_{t-1}}\right)^i\left(1-\frac{k-i+\gamma}{W_{t-1}}\right)^{mX_t^{(1)}-i}|\mathcal{F}_{t-1}\right] + \ldots \\ + p_e^{(r)}\mathbb{E}\left[\binom{mX_t^{(r)}}{i}\left(\frac{k-i+\gamma}{W_{t-1}}\right)^i\left(1-\frac{k-i+\gamma}{W_{t-1}}\right)^{mX_t^{(r)}-i}|\mathcal{F}_{t-1}\right]. \end{aligned} \tag{5.16}$$

Now, for any random variable Z_t with constant expectation μ, independent of the σ-algebra $\mathcal{F}_{t-1}$, and such that $1 \leq Z_t < t^{1/4}$, by Bernoulli's inequality we have

$$\begin{aligned} \mathbb{E}\left[\left(1-\frac{k+\gamma}{W_{t-1}}\right)^{mZ_t}|\mathcal{F}_{t-1}\right] &\geqslant \mathbb{E}\left[\left(1-\frac{mZ_t(k+\gamma)}{W_{t-1}}\right)|\mathcal{F}_{t-1}\right] \\ &= 1-\frac{m\mu(k+\gamma)}{W_{t-1}}. \end{aligned} \tag{5.17}$$

On the other hand (using the fact that for $x \in [0,1]$ and $n \in \mathbb{N}$ we have $(1-x)^n \leq \frac{1}{1+nx}$):

$$\begin{aligned} \mathbb{E}\left[\left(1-\frac{k+\gamma}{W_{t-1}}\right)^{mZ_t}|\mathcal{F}_{t-1}\right] &\leqslant \mathbb{E}\left[\frac{1}{1+\frac{mZ_t(k+\gamma)}{W_{t-1}}}|\mathcal{F}_{t-1}\right] \\ &= \mathbb{E}\left[1-\frac{mZ_t(k+\gamma)}{W_{t-1}+mZ(k+\gamma)}|\mathcal{F}_{t-1}\right] \\ &\leqslant \mathbb{E}\left[1-\frac{mZ_t(k+\gamma)}{W_{t-1}}+\frac{(mZ_t(k+\gamma))^2}{W_{t-1}^2}|\mathcal{F}_{t-1}\right] \\ &\leqslant 1-\frac{m\mu(k+\gamma)}{W_{t-1}}+\frac{t^{1/2}(m(k+\gamma))^2}{W_{t-1}^2}, \end{aligned} \tag{5.18}$$

where the last inequality follows from the assumption $Z_t < t^{1/4}$. Now, let us consider the master equation (5.14) for $\mathbb{E}[N_{k,t}]$ term by term. We start with the expected number of vertices that had degree k at time $t-1$ and are still of degree k at time t. By (5.17), Lemma 14 and the fact that $N_{k,t-1} \leq t$ we get

$$\begin{aligned} \mathbb{E}[N_{k,t-1}Q_{0,k,t}] &\geqslant \mathbb{E}\left[N_{k,t-1}\left(1-\frac{(k+\gamma)m(p_{ve}(\mu_0-1)+p_e^{(1)}\mu_1+\ldots+p_e^{(r)}\mu_r)}{W_{t-1}}\right)\right] \\ &= \mathbb{E}\left[N_{k,t-1}\left(1-\frac{(k+\gamma)(\bar{D}-mp_{ve})}{W_{t-1}}\right)\right] \\ &\geqslant \mathbb{E}[N_{k,t-1}]\left(1-\frac{(k+\gamma)(\bar{D}-mp_{ve})}{\mathbb{E}[W_{t-1}]-mt^{3/4}\sqrt{2\ln t}}\right)-t\cdot\frac{1}{t^4}. \end{aligned}$$

To get the last inequality one needs to conduct calculations analogous to those from

the proof of Lemma 12. By 5.18 and additionally using the fact that $W_{t-1} \geq 1$

$$\begin{aligned}\mathbb{E}[N_{k,t-1}Q_{0,k,t}] &\leqslant \mathbb{E}\left[N_{k,t-1}\left(1 - \frac{(k+\gamma)(\bar{D} - mp_{ve})}{W_{t-1}} + \frac{t^{1/2}(p_{ve}+p_e)(m(k+\gamma))^2}{W_{t-1}^2}\right)\right] \\ &\leqslant \mathbb{E}[N_{k,t-1}]\left(1 - \frac{(k+\gamma)(\bar{D} - mp_{ve})}{\mathbb{E}[W_{t-1}] + mt^{3/4}\sqrt{2\ln t}} + \frac{t^{1/2}(p_{ve}+p_e)(m(k+\gamma))^2}{(\mathbb{E}[W_{t-1}] - mt^{3/4}\sqrt{2\ln t})^2}\right) \\ &\quad + \left(t + t^{3/2}(p_{ve}+p_e)(m(k+\gamma))^2\right)\cdot\frac{1}{t^4}.\end{aligned}$$

Again, for the last inequality, proceed as in the proof of Lemma 12. Since $\mathbb{E}[W_{t-1}] = \bar{D}(t-1) + \gamma\bar{V}(t-1)$ and $\mathbb{E}[N_{k,t-1}] \leqslant t$, we obtain for fixed k

$$\mathbb{E}[N_{k,t-1}Q_{0,k,t}] = \mathbb{E}[N_{k,t-1}]\left(1 - \frac{(k+\gamma)(\bar{D} - mp_{ve})}{t(\bar{D}+\gamma\bar{V}) + \mathcal{O}(t^{3/4}\sqrt{\ln t})}\right) + \mathcal{O}\left(\frac{1}{\sqrt{t}}\right). \quad (5.19)$$

We treat $\mathbb{E}[N_{k-1,t-1}Q_{1,k-1,t}]$ similarly. On one hand we have

$$\begin{aligned}Q_{1,k-1,t} &\geqslant p_{ve}\mathbb{E}\left[m(Y_t-1)\frac{k-1+\gamma}{W_{t-1}}\left(1 - \frac{mY_t(k-1+\gamma)}{W_{t-1}}\right)|\mathcal{F}_{t-1}\right] \\ &\quad + p_e^{(1)}\mathbb{E}\left[mX_t^{(1)}\frac{k-1+\gamma}{W_{t-1}}\left(1 - \frac{mX_t^{(1)}(k-1+\gamma)}{W_{t-1}}\right)|\mathcal{F}_{t-1}\right] + \dots \\ &\quad + p_e^{(r)}\mathbb{E}\left[mX_t^{(r)}\frac{k-1+\gamma}{W_{t-1}}\left(1 - \frac{mX_t^{(r)}(k-1+\gamma)}{W_{t-1}}\right)|\mathcal{F}_{t-1}\right] \\ &\geqslant p_{ve}\mathbb{E}\left[m(Y_t-1)\frac{k-1+\gamma}{W_{t-1}}|\mathcal{F}_{t-1}\right] - p_{ve}\mathbb{E}\left[\frac{Y_t^2(m(k-1+\gamma))^2}{W_{t-1}^2}|\mathcal{F}_{t-1}\right] + \dots \\ &\quad + p_e^{(r)}\mathbb{E}\left[m(X_t^{(r)})\frac{k-1+\gamma}{W_{t-1}}|\mathcal{F}_{t-1}\right] - p_e^{(r)}\mathbb{E}\left[\frac{(X_t^{(r)})^2(m(k-1+\gamma))^2}{W_{t-1}^2}|\mathcal{F}_{t-1}\right] \\ &\geqslant \frac{p_{ve}m(\mu_0-1)(k-1+\gamma)}{W_{t-1}} - \frac{t^{1/2}p_{ve}(m(k-1+\gamma))^2}{W_{t-1}^2} + \dots \\ &\quad + \frac{p_e^{(r)}m\mu_r(k-1+\gamma)}{W_{t-1}} - \frac{t^{1/2}p_e^{(r)}(m(k-1+\gamma))^2}{W_{t-1}^2} \\ &= \frac{(k-1+\gamma)(\bar{D} - mp_{ve})}{W_{t-1}} - \frac{t^{1/2}(p_{ve}+p_e)(m(k-1+\gamma))^2}{W_{t-1}^2}\end{aligned}$$

(the last inequality follows from assumptions $Y_t < t^{1/4}$ and $X_t^{(i)} < t^{1/4}$), while on the other hand

$$\begin{aligned}Q_{1,k-1,t} &\leqslant p_{ve}\mathbb{E}\left[m(Y_t-1)\frac{k-1+\gamma}{W_{t-1}}|\mathcal{F}_{t-1}\right] + \dots + p_e^{(r)}\mathbb{E}\left[mX_t^{(r)}\frac{k-1+\gamma}{W_{t-1}}|\mathcal{F}_{t-1}\right] \\ &\leqslant \frac{(k-1+\gamma)(\bar{D} - mp_{ve})}{W_{t-1}}.\end{aligned} \quad (5.20)$$

Again, by Lemma 14, the fact that $N_{k-1,t-1} \leqslant t$ and $N_{k-1,t-1}/W_{t-1} \leqslant 1$ for fixed k we get

$$\mathbb{E}[N_{k-1,t-1}Q_{1,k-1,t}] = \mathbb{E}[N_{k-1,t-1}]\left(\frac{(k-1+\gamma)(\bar{D} - mp_{ve})}{t(\bar{D}+\gamma\bar{V}) + \mathcal{O}(t^{3/4}\sqrt{\ln t})}\right) + \mathcal{O}\left(\frac{1}{\sqrt{t}}\right). \quad (5.21)$$

The terms from equations (5.19) and (5.21) are those dominating in master equation (5.14). For the sum of other terms we have the following upper bound when k is fixed (the fourth inequality follows from upper bounding the sums by infinite geometric series and the asymptotics in the last line follows from Lemma 14)

$$
\begin{aligned}
\sum_{i=2}^{\min\{k,mZ_t\}} \mathbb{E}[N_{k-i,t-1}Q_{i,k-i,t}] &\leqslant t\cdot\sum_{i=2}^{k}\mathbb{E}[Q_{i,k-i,t}] \\
&\leqslant t\cdot\mathbb{E}\left[\sum_{i=2}^{k}\left(p_{ve}\mathbb{E}\left[\binom{m(Y_t-1)}{i}\left(\frac{k-i+\gamma}{W_{t-1}}\right)^i|\mathcal{F}_{t-1}\right]\right.\right. \\
&\qquad + p_e^{(1)}\mathbb{E}\left[\binom{mX_t^{(1)}}{i}\left(\frac{k-i+\gamma}{W_{t-1}}\right)^i|\mathcal{F}_{t-1}\right]+\dots \\
&\qquad \left.\left. +p_e^{(r)}\mathbb{E}\left[\binom{mX_t^{(r)}}{i}\left(\frac{k-i+\gamma}{W_{t-1}}\right)^i|\mathcal{F}_{t-1}\right]\right)\right] \\
&\leqslant t\cdot\mathbb{E}\left[\mathbb{E}\left[\sum_{i=2}^{k}\left(p_{ve}(mY_t)^i\left(\frac{k+\gamma}{W_{t-1}}\right)^i+\dots\right.\right.\right. \\
&\qquad \left.\left.\left. +p_e^{(r)}(mX_t^{(r)})^i\left(\frac{k+\gamma}{W_{t-1}}\right)^i\right)|\mathcal{F}_{t-1}\right]\right] \\
&\leqslant t\cdot\mathbb{E}\left[\mathbb{E}\left[p_{ve}\frac{(m(k+\gamma)Y_t)^2}{W_{t-1}^2}\frac{1}{1-\frac{m(k+\gamma)Y_t}{W_{t-1}}}+\dots\right.\right. \\
&\qquad \left.\left. +p_e^{(r)}\frac{(m(k+\gamma)X_t^{(r)})^2}{W_{t-1}^2}\frac{1}{1-\frac{m(k+\gamma)X_t^{(r)}}{W_{t-1}}}|\mathcal{F}_{t-1}\right]\right] \\
&\leqslant t\cdot\mathbb{E}\left[p_{ve}\frac{(m(k+\gamma)t^{1/4})^2}{W_{t-1}^2}\frac{1}{1-\frac{m(k+\gamma)t^{1/4}}{W_{t-1}}}+\dots\right. \\
&\qquad \left. +p_e^{(r)}\frac{(m(k+\gamma)t^{1/4})^2}{W_{t-1}^2}\frac{1}{1-\frac{m(k+\gamma)t^{1/4}}{W_{t-1}}}\right] \\
&= t\cdot\mathbb{E}\left[\frac{(p_{ve}+p_e)(m(k+\gamma))^2t^{1/2}}{W_{t-1}^2}\frac{W_{t-1}}{W_{t-1}-m(k+\gamma)t^{1/4}}\right] \\
&= (p_{ve}+p_e)(m(k+\gamma))^2t^{3/2}\cdot\mathbb{E}\left[\frac{1}{W_{t-1}(W_{t-1}-m(k+\gamma)t^{1/4})}\right] \\
&\sim (p_{ve}+p_e)(m(k+\gamma))^2t^{3/2}\cdot\frac{1}{t^2}=\mathcal{O}\left(\frac{1}{\sqrt{t}}\right).
\end{aligned}
\tag{5.22}
$$

Plugging 5.19, 5.21 and 5.22 into master equation (5.13) and (5.14) we obtain

$$
\mathbb{E}[N_{0,t}] = \mathbb{E}[N_{0,t-1}]\left(1-\frac{\gamma(\bar{D}-mp_{ve})}{t(\bar{D}+\gamma\bar{V})+\mathcal{O}(t^{3/4}\sqrt{\ln t})}\right)+p_v+\mathcal{O}\left(\frac{1}{\sqrt{t}}\right) \tag{5.23}
$$

and

$$\mathbb{E}[N_{k,t}] = \mathbb{E}[N_{k,t-1}]\left(1 - \frac{(k+\gamma)(\bar{D} - mp_{ve})}{t(\bar{D} + \gamma\bar{V}) + \mathcal{O}(t^{3/4}\sqrt{\ln t})}\right) + \mathbb{E}[N_{k-1,t-1}]\left(\frac{(k-1+\gamma)(\bar{D} - mp_{ve})}{t(\bar{D} + \gamma\bar{V}) + \mathcal{O}(t^{3/4}\sqrt{\ln t})}\right) + \delta_{k,m}p_{ve} + \mathcal{O}\left(\frac{1}{\sqrt{t}}\right). \tag{5.24}$$

For $k \geq 0$ by L_k denote the limit

$$L_k = \lim_{t\to\infty} \frac{\mathbb{E}[N_{k,t}]}{t}.$$

First we prove that the limit L_0 exists. We apply Lemma 15 to equation (5.23) by setting

$$b_t = \frac{\gamma(\bar{D} - mp_{ve})}{\bar{D} + \gamma\bar{V} + \mathcal{O}(t^{3/4}\sqrt{\ln t}/t)} \quad \text{and} \quad c_t = p_v + \mathcal{O}\left(\frac{1}{\sqrt{t}}\right).$$

We get

$$\lim_{t\to\infty} b_t = \frac{\gamma(\bar{D} - mp_{ve})}{\bar{D} + \gamma\bar{V}} \qquad \text{and} \qquad \lim_{t\to\infty} c_t = p_v,$$

therefore

$$L_0 = \frac{p_v}{1 + \frac{\gamma(\bar{D} - mp_{ve})}{\bar{D} + \gamma\bar{V}}} = \frac{p_v \frac{\bar{D} + \gamma\bar{V}}{\bar{D} - mp_{ve}}}{\frac{\bar{D} + \gamma\bar{V}}{\bar{D} - mp_{ve}} + \gamma}.$$

Now, we assume that the limit L_{k-1} exists and we will show by induction on k that L_k exists. Again, applying Lemma 15 to equation (5.24) with

$$b_t = \frac{(k+\gamma)(\bar{D} - mp_{ve})}{\bar{D} + \gamma\bar{V} + \mathcal{O}(t^{3/4}\sqrt{\ln t}/t)}$$

and

$$c_t = \frac{\mathbb{E}[N_{k-1,t-1}]}{t}\left(\frac{(k-1+\gamma)(\bar{D} - mp_{ve})}{\bar{D} + \gamma\bar{V} + \mathcal{O}(t^{3/4}\sqrt{\ln t}/t)}\right) + \delta_{k,m}p_{ve} + \mathcal{O}\left(\frac{1}{\sqrt{t}}\right)$$

we get

$$\lim_{t\to\infty} b_t = \frac{(k+\gamma)(\bar{D} - mp_{ve})}{\bar{D} + \gamma\bar{V}}$$

and

$$\lim_{t\to\infty} c_t = L_{k-1}\frac{(k-1+\gamma)(\bar{D} - mp_{ve})}{\bar{D} + \gamma\bar{V}} + \delta_{k,m}p_{ve},$$

therefore

$$L_k = \frac{L_{k-1}\frac{(k-1+\gamma)(\bar{D} - mp_{ve})}{\bar{D} + \gamma\bar{V}} + \delta_{k,m}p_{ve}}{1 + \frac{(k+\gamma)(\bar{D} - mp_{ve})}{\bar{D} + \gamma\bar{V}}} = \frac{L_{k-1}(k-1+\gamma) + \delta_{k,m}p_{ve}\frac{\bar{D} + \gamma\bar{V}}{\bar{D} - mp_{ve}}}{k + \gamma + \frac{\bar{D} + \gamma\bar{V}}{\bar{D} - mp_{ve}}}. \tag{5.25}$$

From now on, for simplicity of notation, we put $D = \frac{\bar{D} + \gamma\bar{V}}{\bar{D} - mp_{ve}}$ thus we have

$$L_0 = \frac{p_v D}{\gamma + D} \qquad \text{and} \qquad L_k = \frac{L_{k-1}(k-1+\gamma) + \delta_{k,m}p_{ve}D}{k + \gamma + D}.$$

When $k \in \{1, 2, \ldots, m-1\}$, iterating over k gives

$$L_k = L_0 \cdot \prod_{\ell=1}^{k} \frac{\ell - 1 + \gamma}{\ell + \gamma + D} = \frac{p_v D}{\gamma + D} \prod_{\ell=1}^{k} \frac{\ell - 1 + \gamma}{\ell + \gamma + D}$$

and when $k \geqslant m$

$$\begin{aligned} L_k &= \frac{p_v D}{\gamma + D} \left(\prod_{\ell=1}^{k} \frac{\ell - 1 + \gamma}{\ell + \gamma + D} \right) + \frac{p_{ve} D}{m + \gamma + D} \left(\prod_{\ell=m+1}^{k} \frac{\ell - 1 + \gamma}{\ell + \gamma + D} \right) \\ &= \left(\frac{p_v D}{\gamma + D} \left(\prod_{\ell=1}^{m} \frac{\ell - 1 + \gamma}{\ell + \gamma + D} \right) + \frac{p_{ve} D}{m + \gamma + D} \right) \left(\prod_{\ell=m+1}^{k} \frac{\ell - 1 + \gamma}{\ell + \gamma + D} \right) \\ &= \left(\frac{p_v D}{\gamma + D} \frac{\Gamma(m + \gamma)}{\Gamma(\gamma)} \frac{\Gamma(\gamma + D + 1)}{\Gamma(m + \gamma + D + 1)} + \frac{p_{ve} D}{m + \gamma + D} \right) \\ &\quad \cdot \frac{\Gamma(m + \gamma + D + 1)}{\Gamma(m + \gamma)} \frac{\Gamma(k + \gamma)}{\Gamma(k + \gamma + D + 1)}, \end{aligned}$$

where $\Gamma(x)$ is the gamma function. Since $\lim_{k\to\infty} \frac{\Gamma(k) k^\alpha}{\Gamma(k+\alpha)} = 1$ for constant $\alpha \in \mathbb{R}$, we get

$$\lim_{t\to\infty} \frac{\mathbb{E}[N_{k,t}]}{t} = L_k \sim c \cdot k^{-(1+D)}$$

("$\sim$" refers to the limit by $k \to \infty$) for

$$c = p_v D \cdot \frac{\Gamma(\gamma + D)}{\Gamma(\gamma)} + p_{ve} D \cdot \frac{\Gamma(m + \gamma + D)}{\Gamma(m + \gamma)}.$$

Hence, by Lemma 12, we obtain

$$\lim_{t\to\infty} \mathbb{E}\left[\frac{N_{k,t}}{|V_t|} \right] \sim \frac{c}{p_v + p_{ve}} k^{-(1+D)}.$$

We infer that the degree distribution of our hypergraph follows power-law with

$$\beta = 1 + D = 1 + \frac{\bar{D} + \gamma \bar{V}}{\bar{D} - m p_{ve}} = 2 + \frac{\gamma \bar{V} + m p_{ve}}{\bar{D} - m p_{ve}}.$$

□

Below we present a bunch of examples showing that our theorem generalises the results for the degree distribution of well known models.

Example 1 (Barabási-Albert graph model, [2])**.** *In a single time step we always add one new vertex and attach it with m edges (in proportion to degrees) to existing structure. Thus $p_v = 0$, $p_{ve} = 1$, $p_e = 0$, $\bar{V} = 1$, $Y_t = 2$, $\bar{D} = 2m$, $\gamma = 0$ and we get $\beta = 2 + \frac{m}{2m-m} = 3$.*

Example 2 (Chung-Lu graph model, [9])**.** *In a single time step: we either (with probability p) add one new vertex and attach it with an edge (in proportion to degrees) to existing structure; otherwise we just add an edge (in proportion to degrees) to existing structure. Thus $p_v = 0$, $p_{ve} = p$, $p_e = 1-p$, $\bar{V} = p$, $Y_t = 2$, $r = 1$, $X_t^{(1)} = 2$, $\bar{D} = 2$, $m = 1$, $\gamma = 0$ and we get $\beta = 2 + \frac{p}{2-p}$.*

Example 3 (Avin et al. hypergraph model, [1])**.** *In a single time step we either (with probability p) add one new vertex and attach it with a hyperedge of cardinality Y_t (in proportion to degrees) to existing structure; otherwise we just add a hyperedge of cardinality Y_t to existing structure. The assumptions on Y_t and the sum of degrees D_t are: 1.* $\lim_{t\to\infty} \frac{\mathbb{E}[D_{t-1}]/t}{\mathbb{E}[Y_t]-p_{ve}} = D \in (0,\infty)$*, 2.* $\mathbb{E}[|\frac{1}{D_t} - \frac{1}{\mathbb{E}[D_t]}|] = o(1/t)$*, 3.* $\mathbb{E}\left[\frac{Y_t^2}{D_{t-1}^2}\right] = o(1/t)$*. The result from [1] states that the degree distribution of the resulting hypergraph follows a power-law with $\beta = 1 + D$. Note that in our model* $\lim_{t\to\infty} \frac{\mathbb{E}[D_{t-1}]/t}{\mathbb{E}[Y_t]-p_{ve}} = \frac{\bar{D}}{\bar{D}-p_{ve}}$*. Setting $p_v = 0$, $p_{ve} = p$, $p_e = 1-p$, $\bar{V} = p$, $m = 1$, $\gamma = 0$ we get* $\beta = 2 + \frac{p_{ve}}{\bar{D}-p_{ve}} = 1 + \frac{\bar{D}}{\bar{D}-p_{ve}} = 1 + D$*.*

Remark 6. *Even though our result from this section may seem similar to what was obtained by Avin et al., it is easy to indicate cases that are covered by our model but not by the one from [1] and vice versa. Indeed, the model from [1] admits a wide range of distributions for Y_t. In particular, as authors underline, three mentioned assumptions hold for Y_t which is polynomial in t. This is the case not covered by our model (we upper bound Y_t by $t^{1/4}$) but we also can not think of real-life examples that would require bigger hyperedges. Whereas we can think of some natural examples that break requirements from [1] but are admissible in our model. Put $Y_t = 2$ if t is odd and $Y_t = 3$ if t is even. Then* $\lim\limits_{\substack{t\to\infty \\ t \text{ - even}}} \frac{\mathbb{E}[D_{t-1}]/t}{\mathbb{E}[Y_t]-p_{ve}} = \frac{5/2}{3-p_{ve}}$ *and* $\lim\limits_{\substack{t\to\infty \\ t \text{ - odd}}} \frac{\mathbb{E}[D_{t-1}]/t}{\mathbb{E}[Y_t]-p_{ve}} = \frac{5/2}{2-p_{ve}}$ *thus the limit* $\lim_{t\to\infty} \frac{\mathbb{E}[D_{t-1}]/t}{\mathbb{E}[Y_t]-p_{ve}}$ *does not exist. Whereas in our model we are allowed to put $r = 2$,* $p_e^{(1)} = p_e^{(2)} = 1/2$*,* $X_t^{(1)} = 2$*,* $X_t^{(2)} = 3$ *which probabilistically simulates stated example.*

5.4 Hypergraph model with high modularity

In this section we present a new preferential attachment hypergraph model which features partition into communities. To the best of our knowledge no mathematical model so far consolidated preferential attachment, possibility of having hyperedges and clear community structure. We prove that its degree distribution follows a power-law in Section 5.6, and study its modularity in Section 5.7.

We denote our hypergraph by $G_t = (V_t, E_t)$. At each time step either a new vertex (*vertex-step*) or a new hyperedge (*hyperedge-step*) is added to the existing structure. The set of vertices of G_t is partitioned into r communities $V_t = C_t^{(1)} \dot\cup C_t^{(2)} \dot\cup \ldots \dot\cup C_t^{(r)}$. Whenever a new vertex is added to G_t it is assigned to the one of r communities and stays there forever.

Hypergraph model G is characterized by six parameters:

1. G_0 - initial hypergraph seen at time $t = 0$ with vertices partitioned into r communities $V_0 = C_0^{(1)} \dot\cup C_0^{(2)} \dot\cup \ldots \dot\cup C_0^{(r)}$;
2. $p \in (0,1)$ - the probability of taking a vertex-step;
3. vector $M = (m_1, m_2, \ldots, m_r)$ with all m_i positive, constant and summing up to 1; m_i is the probability that a randomly chosen vertex belongs to $C_t^{(i)}$;
4. d-dimensional matrix $P_{r\times\ldots\times r}$ of hyperedge probabilities ($P_{i_1,i_2,\ldots,i_d}$ is the probability that communities $i_1, \ldots, i_d$ share a hyperedge); d is the upper bound for the number of communities shared by a single hyperedge;

5. $X = ((X_0^{(1)}, X_1^{(1)}, \ldots), (X_0^{(2)}, X_1^{(2)}, \ldots), \ldots, (X_0^{(d)}, X_1^{(d)}, \ldots))$ - d sequences of independent random variables indicating the number of vertices from a particular community involved in a newly created hyperedge;

6. $\gamma \geqslant 0$ - parameter appearing in the formula for the probability of choosing a particular vertex to a newly created hyperedge.

We build a structure of $G(G_0, p, M, X, P, \gamma)$ starting with some initial hypergraph G_0. Here G_0 consists of r disjoint hyperedges of cardinality 1. All vertices are assigned to different communities. 'Vertices are chosen from $C_t^{(i)}$ in proportion to degrees' means that vertices are chosen independently (possibly with repetitions) and the probability that any u from $C_t^{(i)}$ is chosen equals

$$\mathbb{P}[u \text{ is chosen}] = \frac{\deg_t(u) + \gamma}{\sum_{v \in C_t^{(i)}} (\deg_t(v) + \gamma)},$$

($\deg_t(v)$ is the degree of v in G_t). For $t \geqslant 0$, G_{t+1} is obtained from G_t as follows:

- With probability p add one new isolated vertex and assign it to one of r communities according to a categorical distribution given by vector M.
- Otherwise, create a hyperedge:
  - according to P select N communities (N is a random variable depending on P) that will share a hyperedge being created, say $C_t^{(i_1)}, C_t^{(i_2)}, \ldots, C_t^{(i_N)}$;
  - assign selected communities to N random variables chosen from $\{X_t^{(1)}, \ldots, X_t^{(r)}\}$ uniformly independently at random, say to $X_t^{(j_1)}, \ldots, X_t^{(j_N)}$;
  - for each $s \in \{1, \ldots, N\}$ select $X_t^{(j_s)}$ vertices from $C_t^{(i_s)}$ in proportion to degrees;
  - create a hyperedge consisting of all selected vertices.

5.5 Study of the co-authorship hypergraph

Before showing the power-law degree distribution and the strong modularity of the model presented in Section 5.4, we first study the co-publication hypergraph extracted from the Scopus database. The study of this graph allows us to determine the values of the model parameters (p, M, X, P, γ) in the real co-publication graph.

5.5.1 Presentation of the network

We first present the co-publication hypergraph. This work is a collaboration with economists from the laboratory of GREDEG and of the SKEMA Business School. The overall goal is to study the impact of research fundings on productivity and pluridisciplinarity. In that purpose, they crawled publication metadata in Scopus, a transdisciplinary database of abstracts and citations of scientific publications of the publisher Elsevier. The extraction contains all papers published from 1990 to 2018 with at least one French co-author, leading to 3.9 million papers from 2.2 million distinct authors from various disciplines such as ecology, computer science,

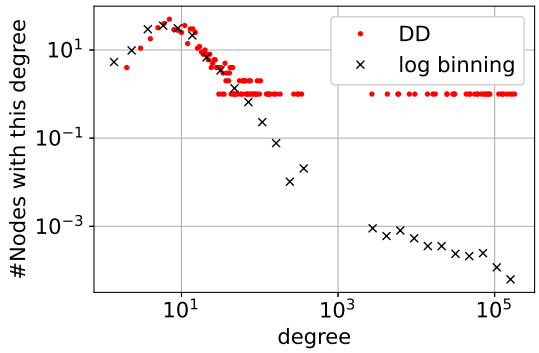


Figure 5.1: Distribution of the sizes of the communities for the co-publication hypergraph.

music, philosophy and so on. Note that among those papers, we only kept the ones with fewer than 100 authors. We believe this arbitrary threshold is reasonable in order to avoid papers which would lead to high hyperedges with no real collaboration meaning. We cleaned those raw data in order to get rid of some singularities (papers without authors, ...) and kept only for each paper its id, its publication date, and the list of its authors.

We then represent those data into a hypergraph: a node corresponds to an author, and a hyperedge corresponds to a paper written by a set of authors - e.g., a paper with the three authors Alice, Bob, and Clara will give a hyperedge between the three nodes associated to Alice, Bob, and Clara.

5.5.2 Study of the properties

We now compute the parameters (p, M, X, P, γ) in the co-publication hypergraph. Those parameters will be used in Section 5.7 to compare the co-publication hypergraph with randomly generated hypergraphs from the model with same parameters.

Largest Connected Component

We first study the connected components. The largest connected component contains the main part of the graph, with 94.22% of the nodes and 99.23% of the edges. From now on, we thus work on this largest connected component.

Number of communities

In order to study the communities, our first idea was to use an implementation for community detection in hypergraphs in the Julia language [23]. This implementation uses the modularity for hypergraphs from Definition 4 introduced in [17]. To the best of our knowledge, this is the only implementation available nowadays to find communities directly using hypergraphs properties. However, if this implementation is useful for smaller graphs, the computation on the co-publication graph with millions of nodes and hyperedges was really long to run and returned bad-quality partitions of the nodes, resulting into a small value of modularity.

Another way to partition the nodes into communities is to use a projection of the hypergraph into a weighted graph, and use well-known partitioning algorithms

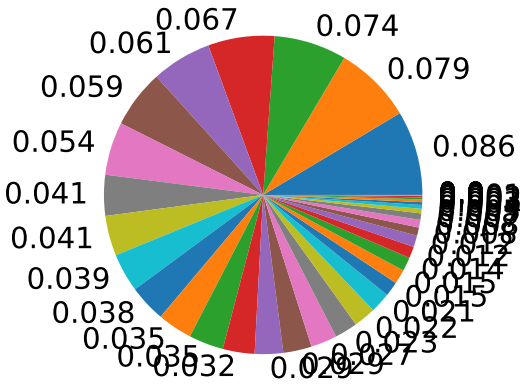


Figure 5.2: Proportion of nodes contained in communities of size bigger than 100.

on this graph. Note that, since the partition into communities only concerns nodes and not hyperedges, the same partition can be used to compute both the modularity of the projected graph and the one of the hypergraph.

The projected graph is built such that each hyperedge is transformed into a clique with edges of weight $1/k$, with k the size of the hyperedge. Edges are weighted in order to moderate the influence of the size of the hyperedges. Note that this chosen weight is an arbitrary choice, and the study of other choices (such that taking weights 1 or $2/(k(k-1))$) would be interesting for future works. However, as we will see, this projection already gives good results for partitioning. We also remind that the projected graph is only used to find the partition, while the hypergraph is used for the rest of the study.

To find a good partition for the nodes, we use the Leiden's algorithm [24], known to give faster and better results than the well-known Louvain's algorithm [3]. We also used the Louvain's algorithm in order to compare both methods, and confirm that, in our case, the Leiden's algorithm gave better results and in a smaller amount of time. We apply the Leiden's algorithm in the projected graph. We obtain 595 communities. Using those communities to compute the modularity on the co-publication hypergraph using Definition 4 gives a value of 0.63. This high value confirmed that the obtained partition reflects well the hypergraph community structure. Figure 5.1 presents the distribution of the obtained community sizes.

For the rest of the study, we decided to keep only communities with more than one hundred nodes, in order to get rid of the communities too small to have a real meaning. Note that those small communities only concern 0.44% of the nodes. Doing so leaves us with a total of 47 communities. For those 47 communities, Figure 5.2 represents the proportions of the sizes of the kept communities. Note that this corresponds to the parameter M of the model presented in Section 5.4.

Number of communities involved by a hyperedge

Having a partition for the communities, we now study how many communities each hyperedge involves. Figure 5.3 presents the proportion of hyperedges in which exactly X communities are involved. For instance, if you take a hyperedge randomly, you have 74.4% chance that all its nodes belong to the same community. In particular, we see that 93.9% of hyperedges are contained into one or two communities.

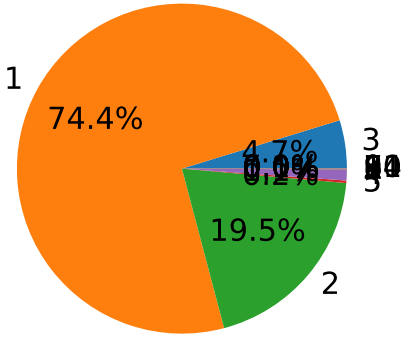


Figure 5.3: Number of communities implied in a hyperedge.

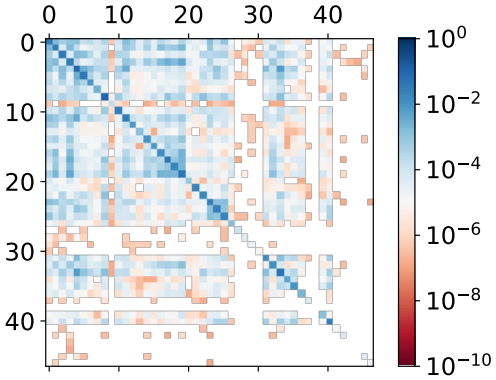


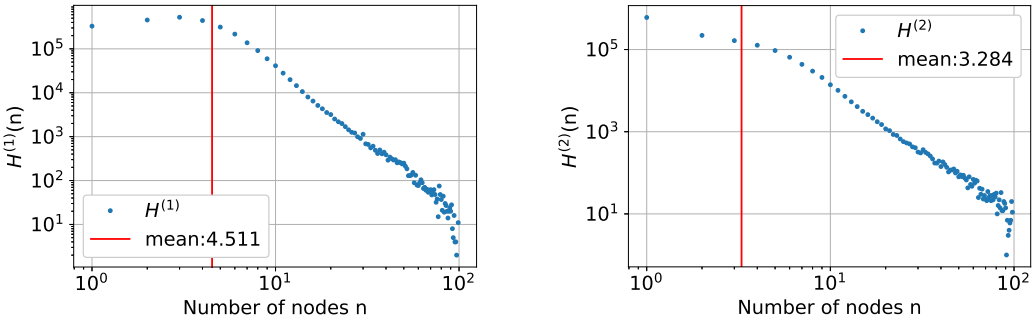
Figure 5.4: Matrix P: $P_{i,j}$ corresponds to the probability that a hyperedge contains nodes of communities i and j.

This justifies the choice of a 2-dimensional matrix P in the application of the model of Section 5.4.

Real values for the parameters of the model

Finally, we computed the different values for the parameters of the model presented in Section 5.4:

- $p = \frac{\text{number of nodes}}{\text{number of nodes+number of edges}} = 0.3628$;
- M is given by Figure 5.2;
- The 2 distributions for picking the sequences of random variables X are represented in Figure 5.5. Figure 5.5a (resp. Figure 5.5b) presents the distribution of the number of papers with a given number of authors and for which authors all belong to the same community (resp. belong to exactly two communities);
- P is presented in Figure 5.4; in particular, we note that the non-diagonal elements sums to $\alpha = 0.208$;
- γ is chosen arbitrarily.



(a) $H^{(1)}(n)$. (b) $H^{(2)}(n)$.

Figure 5.5: $H^{(r)}(n)$: distribution of the number of papers with n authors who belong to exactly r communities.

5.6 Degree distribution of $\mathbf{G(G_0, p, M, X, P, \gamma)}$

We show in this section that a graph G built with the model presented in Section 5.4 have degree distributions following a power-law. A power-law degree distribution of G comes from the fact that each community of G behaves over time as the hypergraph model H presented in Section 5.3.1. Thus the degree distribution of each community follows a power-law. This we prove using directly Theorem 5.

Theorem 6. *Consider a hypergraph $G = G(G_0, p, M, X, P, \gamma)$ for all $t > 0$. Let $\mathbb{E}[X_t^{(i)}] = \mu_i$ and $1 \leqslant X_t^{(i)} < t^{1/4}$ for $i \in \{0, 1, \ldots, r\}$. Then the degree distribution of G follows a power-law with $\beta = 2 + \gamma \cdot \min_{j \in \{1,\ldots,r\}} \{\bar{V}_j / \bar{D}_j\}$, where $\bar{V}_j$ is the expected number of vertices added to $C_t^{(j)}$ at a single time step and $\bar{D}_j$ is the expected number of vertices from $C_t^{(j)}$ that increase their degree at a single time step. I.e.,*

$$\beta = 2 + \frac{\gamma p}{(1-p)\frac{\mu_1 + \ldots + \mu_r}{r}} \cdot \min_{j \in \{1,\ldots,r\}} \left\{ \frac{m_j}{s_j} \right\},$$

where s_j is the probability that by creating a new hyperedge a community j is chosen as the one sharing it.

Remark 7. *The value s_j can be derived from P; it is the sum of probabilities of creating a hyperedge between $C^{(j)}$ and any other subset of communities.*

The number of vertices in G_t is a random variable satisfying $|V_t| \sim B(t, p) + r$, while for the number of hyperedges in G_t we have $|E_t| \sim B(t, 1-p) + r$. Note that since $|V_t|$ follows a binomial distribution, Lemma 12 holds also in case of G_t if we replace $p_v + p_{ve}$ by p.

Recall that $N_{k,t}$ stands for the number of vertices in G_t of degree k. For $i \in \{1, 2, \ldots, r\}$ by $N_{k,t}^{(i)}$ we denote the number of vertices of degree k in G_t belonging to community $C_t^{(i)}$. Thus $N_{k,t} = \sum_{i=1}^{r} N_{k,t}^{(i)}$.

Lemma 16. *Consider a single community $C_t^{(j)}$ of a hypergraph G_t. Let $\mathbb{E}[X_t^{(i)}] = \mu_i$ and $1 \leqslant X_t^{(i)} < t^{1/4}$ for $i \in \{0, 1, \ldots, r\}$. Then the degree distribution of vertices from $C_t^{(j)}$ follow a power-law with*

$$\beta_j = 2 + \frac{\gamma \bar{V}_j}{\bar{D}_j}$$

where $\bar{V}_j$ is the expected number of vertices added to $C_t^{(j)}$ at a single time step and $\bar{D}_j$ is the average number of vertices from $C_t^{(j)}$ that increase their degree at a single time step, thus $\bar{V}_j = pm_j$ and $\bar{D}_j = (1-p)s_j\frac{\mu_1+\ldots+\mu_r}{r}$, where s_j is the probability that by creating a new hyperedge a community j is chosen as the one sharing it (we obtain s_j from matrix P - see remark below).

Proof. Note that the community $C_{t+1}^{(j)}$ arises from community $C_t^{(j)}$ choosing at time t only one of the following events according to p, M and P.

- With probability pm_j: Add one new isolated vertex.
- With probability $\frac{(1-p)s_j}{r}$: Select $X_t^{(1)}$ vertices from $C_t^{(j)}$ in proportion to their degrees; these are vertices included in a newly created hyperedge, thus their degrees will increase.
- ...
- With probability $\frac{(1-p)s_j}{r}$: Select $X_t^{(r)}$ vertices from $C_t^{(j)}$ in proportion to their degrees; these are vertices included in a newly created hyperedge, thus their degrees will increase.
- With probability $1-(pm_j+(1-p)s_j)$: Do nothing.

Now, apply Theorem 5 with $p_v = pm_j$, $p_{ve} = 0$, $p_e^{(1)} = p_e^{(2)} = \ldots = p_e^{(r)} = \frac{(1-p)s_j}{r}$ and $m = 1$. We get that the degree distribution of vertices from $C_t^{(j)}$ follow a power-law with

$$\beta_j = 2 + \frac{\gamma\bar{V}_j}{\bar{D}_j} = 2 + \frac{\gamma p m_j}{(1-p)s_j\frac{\mu_1+\ldots+\mu_r}{r}}.$$

□

Theorem 6. We need to prove that $\lim_{t\to\infty}\mathbb{E}\left[\frac{N_{k,t}}{|V_t|}\right] \sim \tilde{c}k^{-\beta}$ for some constant $\tilde{c}$ and β as in the statement of theorem. By Lemma 12 we know that it suffices to show $\lim_{t\to\infty}\frac{\mathbb{E}[N_{k,t}]}{t} \sim ck^{-\beta}$ for some constant c. By Lemma 12 and Lemma 16 we write

$$\begin{aligned}\lim_{t\to\infty}\frac{\mathbb{E}[N_{k,t}]}{t} &= \lim_{t\to\infty}\frac{\mathbb{E}[N_{k,t}^{(1)}]}{t} + \lim_{t\to\infty}\frac{\mathbb{E}[N_{k,t}^{(2)}]}{t} + \ldots + \lim_{t\to\infty}\frac{\mathbb{E}[N_{k,t}^{(r)}]}{t}\\ &\sim c_1k^{-\beta_1} + c_2k^{-\beta_2} + \ldots + c_rk^{-\beta_r}\end{aligned}$$

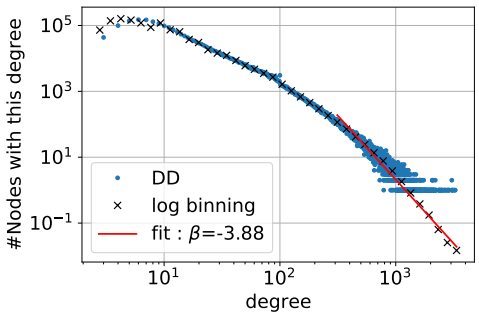
for some constants $c_1,\ldots,c_r$ and $\beta_j = 2+\frac{\gamma\bar{V}_j}{\bar{D}_j}$. Thus $\lim_{t\to\infty}\frac{\mathbb{E}[N_{k,t}]}{t} \sim ck^{-\beta}$, where

$$\beta = \min_{j\in\{1,\ldots,r\}}\{\beta_j\} = 2+\gamma\cdot\min_{j\in\{1,\ldots,r\}}\left\{\frac{\bar{V}_j}{\bar{D}_j}\right\} = 2+\frac{\gamma p}{(1-p)\frac{\mu_1+\ldots+\mu_r}{r}}\cdot\min_{j\in\{1,\ldots,r\}}\left\{\frac{m_j}{s_j}\right\}.$$

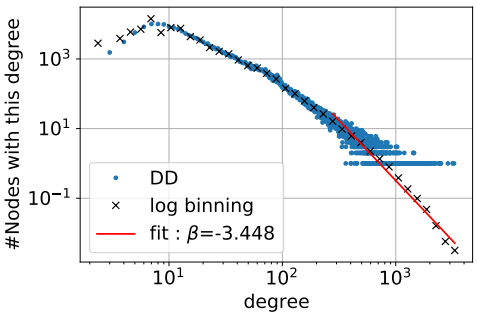
□

In Figure 5.6 we present log-log plots of a power-law distribution fitted to the degree distribution (DD) of the real-life co-authorship hypergraph, denoted by R. R is the same as in Section 5.5. Left chart presents the degree distribution of the whole R while the right one refers only to the biggest community of R found by Leiden algorithm. We see that both those distributions can be considered in first approximation as following a power-law in their tail. Note that this assumption has to be rigorously checked, study that we keep for future works.

Let us also make one remark about the implementation of matrix P.



Graph R



The biggest community of R

Figure 5.6: A power-law distribution fitted to the degree distributions.

Remark 8. *Observe that storing hyperedge probabilities in d-dimensional matrix P we use much more space than we actually should. The same probabilities may repeat many times in P. E.g., when $d = 2$ we get 2-dimensional symmetric matrix P such that $\sum_{i=1}^{r}\sum_{j=1}^{i} p_{ij} = 1$ and the the probability of creating hyperedge between two distinct communities $C^{(i)}$ and $C^{(j)}$ is in matrix P doubled - as p_{ij} and p_{ji}. If we allow for bigger hyperedges it may be repeated much more times. In fact we need to store at most $2^r - 1$ different probabilities (one for each nonempty subset of the set of communities) while in P we store d^r values (in particular, if $d = r$ we store r^r instead $2^r - 1$ values). Nevertheless, for formal proofs this notation is convenient thus we use it at the same time underlining that implementation may be done much more space efficiently.*

5.7 Modularity of $\mathbf{G(G_0, p, M, X, P, \gamma)}$

In this section we give lower bounds for the modularity of $G = G(G_0, p, M, X, P, \gamma)$ in terms of the values from matrix P. We present experimental results showing the advantage in modularity of our model over the one in [1].

5.7.1 Theoretical results

We analyse $G(G_0, p, M, X, P, \gamma) = (V, E)$ obtained up to time t (this time we omit superscripts t). Recall that each vertex from V is assigned to one of r communities, $V = C^{(1)} \dot\cup C^{(2)} \dot\cup \ldots \dot\cup C^{(r)}$. We obtain the lower bound for modularity deriving the modularity score of the partition $\mathcal{C} = \{C^{(1)}, C^{(2)}, \ldots, C^{(r)}\}$. This choice of partition seems obvious provided that matrix P is strongly assortative, i.e., the probabilities of having an edge inside communities are all bigger than the highest probability of having an edge joining different communities. Note that what matters for the value of modularity is the total sum of degrees in each community, not the distribution of degrees. Therefore we do not use the fact that the degree distribution follows a power-law in each community and in the whole model. We just use information from matrix P. Thus, in fact, we derive the lower bound for the modularity of stochastic block model with r communities.

For $\ell \geqslant 1$ $E_\ell \subseteq E$ is the set of hyperedges of cardinality ℓ. First, we state general lower bound for the modularity of G ias a function of matrix P.

Lemma 17. *Let $G = G(G_0, p, M, X, P, \gamma)$ with the size of each hyperedge bounded by d. Let p_i be the probability that a randomly chosen hyperedge is within community $C^{(i)}$ (i.e., all vertices of a hyperedge belong to $C^{(i)}$). By s_i we denote the probability that a randomly chosen hyperedge has at least one vertex in community $C^{(i)}$. Assume also that with high probability $|E_\ell|/|E| \sim a_\ell$ for some constants $a_\ell \in [0,1]$ and $vol(V)/|E| \sim \delta$ for some constant $\delta \in (0, \infty)$. Then whp*

$$\lim_{t\to\infty} q^*(G) \geqslant \sum_{i=1}^{r} p_i - \sum_{i=1}^{r}\sum_{\ell\geqslant 1} a_\ell \left(\frac{(d-1)s_i + p_i}{\delta}\right)^\ell.$$

Proof. Let $\mathcal{C} = \{C^{(1)}, C^{(2)}, \ldots, C^{(r)}\}$. Let also q denote the probability of adding a new hyperedge in a single time step (hence $q = 1 - p$, referring to notation from Section 5.4). Thus with high probability $|E| \sim t \cdot q$ (where '$\sim$' refers to the limit by $t \to \infty$). By Definition 4 we write

$$q^*(G) = \max_{\mathcal{A}} q_{\mathcal{A}}(G) \geqslant q_{\mathcal{C}}(G) = \sum_{i=1}^{r}\left(\frac{|E(C^{(i)})|}{|E|} - \sum_{\ell\geqslant 1}\frac{|E_\ell|}{|E|}\left(\frac{vol(C^{(i)})}{vol(V)}\right)^\ell\right).$$

We obtain that with high probability

$$q_{\mathcal{C}}(G) \sim \sum_{i=1}^{r}\left(\frac{t\cdot q\cdot p_i}{t\cdot q} - \sum_{\ell\geqslant 1} a_\ell\left(\frac{vol(C^{(i)})}{t\cdot q\cdot\delta}\right)^\ell\right).$$

Note that if at a certain time step appears a hyperedge with all vertices contained in $C^{(i)}$, which happens with probability $q \cdot p_i$, it adds up at most d to $vol(C^{(i)})$. If at a certain time step appears a hyperedge joining at least 2 communities with at least one vertex in $C^{(i)}$, which happens with probability $q(s_i - p_i)$, it adds up at most $d-1$ to $vol(C^{(i)})$. Thus we get that with high probability

$$\begin{aligned}\lim_{t\to\infty} q^*(G) &\geqslant \sum_{i=1}^{r} p_i - \sum_{i=1}^{r}\sum_{\ell\geqslant 1} a_\ell\left(\frac{t\cdot q\cdot(dp_i + (d-1)(s_i - p_i))}{t\cdot q\cdot\delta}\right)^\ell \\ &= \sum_{i=1}^{r} p_i - \sum_{i=1}^{r}\sum_{\ell\geqslant 1} a_\ell\left(\frac{(d-1)s_i + p_i}{\delta}\right)^\ell.\end{aligned}$$

□

Remark 9. *Note that for G being 2-uniform (thus simply a graph) this result simplifies significantly to $\lim_{t\to\infty} q^*(G) \geqslant \sum_{i=1}^{r} p_i - 1/4\sum_{i=1}^{r}(s_i + p_i)^2$.*

Below we state the lower bound for the modularity of G in a version in which the knowledge of the whole matrix P is not necessary. Instead we use its two characteristics: α - the probability that a randomly chosen hyperedge joins at least two different communities (may be interpreted as the amount of noise in the network) and β - the maximum value among p_i's for $i \in \{1, 2, \ldots, r\}$. The modularity of the model will be maximised for $\alpha = 0$ (when there are no hyperedges joining different communities) and $\beta = 1/r$ (when all p_i's are equal to $1/r$ thus hyperedges are distributed uniformly across communities).

Lemma 18. *By assumptions from Lemma 17 whp* $\lim_{t\to\infty} q^*(G) \geqslant 1-\alpha-a_1\left(\frac{d}{\delta}\right)((d-2)\alpha+1)-\sum_{\ell\geqslant 2} a_\ell\left(\frac{d}{\delta}\right)^\ell\left((r-1)\beta^\ell+((d-1)\alpha+\beta)^\ell\right)$, *where* $\alpha = 1-\sum_{i=1}^r p_i$ *and* $\beta = \max_{i\in\{1,\ldots,r\}} p_i$.

Proof. Let $\mathcal{C} = \{C^{(1)}, C^{(2)}, \ldots, C^{(r)}\}$ and for $i \in \{1,2,\ldots,r\}$ let $\tilde{s}_i$ be the probability that a randomly chosen hyperedge joins at least two communities and $C^{(i)}$ is one of them. Note that for s_i defined as in Lemma 17 (i.e., the probability that a randomly chosen hyperedge has at least one vertex in $C^{(i)}$) we get $s_i = \tilde{s}_i + p_i$. By Lemma 17 we get that with high probability

$$\begin{aligned}\lim_{t\to\infty} q^*(G) &\geqslant \sum_{i=1}^r p_i - \sum_{i=1}^r\sum_{\ell\geqslant 1} a_\ell\left(\frac{(d-1)\tilde{s}_i+dp_i}{\delta}\right)^\ell \\ &= (1-\alpha) - \sum_{\ell\geqslant 1}\frac{a_\ell}{\delta^\ell}\sum_{i=1}^r((d-1)\tilde{s}_i+dp_i)^\ell \\ &= (1-\alpha) - \frac{a_1}{\delta}\left((d-1)\sum_{i=1}^r\tilde{s}_i + d\sum_{i=1}^r p_i\right) - \sum_{\ell\geqslant 2}\frac{a_\ell}{\delta^\ell}\sum_{i=1}^r((d-1)\tilde{s}_i+dp_i)^\ell.\end{aligned} \tag{5.26}$$

Now, by r_k denote the probability that a randomly chosen hyperedge joins exactly k communities. Note that

$$\sum_{i=1}^r \tilde{s}_i = 2r_2 + 3r_3 + \ldots + dr_d \leqslant d(1-\sum_{i=1}^r p_i) = d\alpha. \tag{5.27}$$

Thus

$$\begin{aligned}\frac{a_1}{\delta}\left((d-1)\sum_{i=1}^r\tilde{s}_i + d\sum_{i=1}^r p_i\right) &\leqslant \frac{a_1}{\delta}\left((d-1)d\alpha + d(1-\alpha)\right) \\ &= a_1\left(\frac{d}{\delta}\right)((d-2)\alpha+1).\end{aligned} \tag{5.28}$$

Moreover,

$$\begin{aligned}\sum_{\ell\geqslant 2}\frac{a_\ell}{\delta^\ell}\sum_{i=1}^r((d-1)\tilde{s}_i+dp_i)^\ell &= \sum_{\ell\geqslant 2}\frac{a_\ell}{\delta^\ell}\sum_{i=1}^r\sum_{k=0}^\ell\binom{\ell}{k}((d-1)\tilde{s}_i)^k(dp_i)^{l-k} \\ &= \sum_{\ell\geqslant 2}\frac{a_\ell}{\delta^\ell}\sum_{k=0}^\ell\binom{\ell}{k}(d-1)^k d^{l-k}\sum_{i=1}^r\tilde{s}_i^k p_i^{l-k} \\ &\leqslant \sum_{\ell\geqslant 2}\frac{a_\ell}{\delta^\ell}\sum_{k=0}^\ell\binom{\ell}{k}(d-1)^k(d\beta)^{l-k}\sum_{i=1}^r\tilde{s}_i^k \\ &= \sum_{\ell\geqslant 2}\frac{a_\ell}{\delta^\ell}\left(r(d\beta)^\ell + \sum_{k=1}^\ell\binom{\ell}{k}(d-1)^k(d\beta)^{l-k}\sum_{i=1}^r\tilde{s}_i^k\right) \\ &\leqslant \sum_{\ell\geqslant 2}\frac{a_\ell}{\delta^\ell}\left(r(d\beta)^\ell + \sum_{k=1}^\ell\binom{\ell}{k}(d-1)^k(d\beta)^{l-k}\left(\sum_{i=1}^r\tilde{s}_i\right)^k\right).\end{aligned}$$

Next, by (5.27) we get

$$\begin{aligned}\sum_{\ell\geqslant 2}\frac{a_\ell}{\delta^\ell}\sum_{i=1}^{r}((d-1)\tilde{s}_i+dp_i)^\ell &\leqslant \sum_{\ell\geqslant 2}\frac{a_\ell}{\delta^\ell}\left(r(d\beta)^\ell+\sum_{k=1}^{\ell}\binom{\ell}{k}(d-1)^k(d\beta)^{l-k}(d\alpha)^k\right)\\ &=\sum_{\ell\geqslant 2}\frac{a_\ell}{\delta^\ell}\left((r-1)(d\beta)^\ell+\sum_{k=0}^{\ell}\binom{\ell}{k}((d-1)d\alpha)^k(d\beta)^{l-k}\right)\\ &=\sum_{\ell\geqslant 2}\frac{a_\ell}{\delta^\ell}\left((r-1)(d\beta)^\ell+((d-1)d\alpha+d\beta)^\ell\right)\\ &=\sum_{\ell\geqslant 2}a_\ell\left(\frac{d}{\delta}\right)^\ell\left((r-1)\beta^\ell+((d-1)\alpha+\beta)^\ell\right).\end{aligned} \tag{5.29}$$

Finally, plugging (5.28) and (5.29) to (5.26) we get that with high probability

$$\begin{aligned}&\lim_{t\to\infty}q^*(G)\geqslant\\ &\quad\geqslant 1-\alpha-a_1\left(\frac{d}{\delta}\right)((d-2)\alpha+1)-\sum_{\ell\geqslant 2}a_\ell\left(\frac{d}{\delta}\right)^\ell\left((r-1)\beta^\ell+((d-1)\alpha+\beta)^\ell\right).\end{aligned}$$

□

Remark 10. *For G being 2-uniform, the result of Lemma 18 simplifies to*

$$\lim_{t\to\infty}q^*(G)\geqslant 1-r\beta^2-\alpha(1+\alpha+2\beta). \tag{5.30}$$

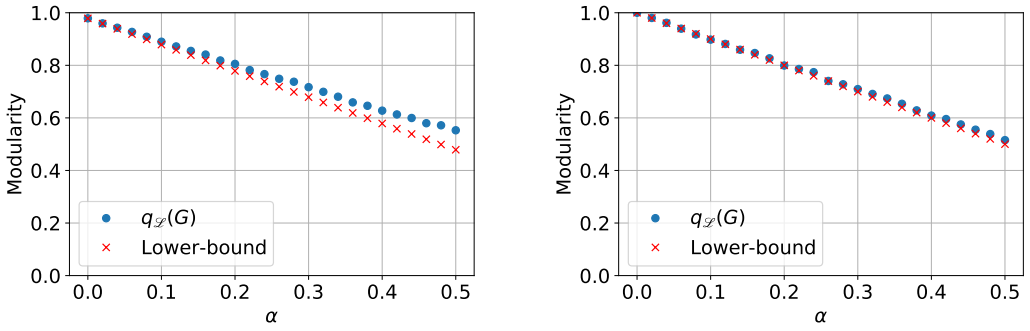
Note that for $\alpha=0$ and $\beta=1/r$, this bound equals $1-1/r$ and is tight, i.e., it is the modularity of the graph with the same number of edges in each of its r communities and no edges between different communities.

Remark 11. *Obtained bounds work well as long as the cardinalities of hyperedges do not differ too much. This is since deriving them we bound the cardinality of each hyperedge by the size of the biggest one. In particular, the bounds are very good in case of uniform hypergraphs - check experimental results below.*

5.7.2 Experimental results

In this subsection we show how the modularity of our model G compares with Avin et al. hypergraph A [1] and with a real-life co-authorship graph R. We also check how good is our theoretical lower bound for modularity.

As discussed in Section 5.5, to get the approximation of modularity of simulated hypergraphs we used Leiden procedure [24] - a popular community detection algorithm for large networks. Calculating modularity is NP-hard [6]. Leiden is nowadays one of the best heuristics trying to find a partition maximising modularity. Therefore we treat its outcome partition as the one whose modularity score is quite precise approximation of the modularity of graphs in question. Every presented modularity score (using Definition 4) refers to a partition returned by Leiden algorithm ran on the flattened hypergraph (i.e., a graph obtained from a hypergraph by exchanging



2-uniform G 20-uniform G

Figure 5.7: Lower bound from Lemma 17 in comparison with the modularity score obtained by Leiden algorithm on simulated uniform hypergraphs G.

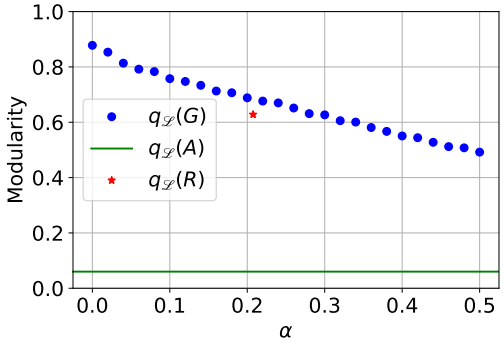


Figure 5.8: Comparison of modularity between our model G, Avin et al. hypergraph A and real co-authorship hypergraph R.

hyperedges with cliques). We did not manage to run Leiden-like algorithm directly for hypergraphs due to their big scale and our technical limitations.

Fig. 5.7 shows the lower bound from Lem. 17 in comparison with the modularity of 2− and 20-uniform hypergraph $G(G_0, p, M, X, P, \gamma)$ on 10^4 vertices, where M is uniform and matrix P has values $(1-\alpha)/47$ (47 is the number of communities also in R) on the diagonal and the rest of probability mass spread uniformly over remaining entries. As we expected - the theoretical bound almost overlapped with the value of modularity in this case.

We then compare the modularity of randomly generated models with the hypergraph R extracted from the Scopus database. We implemented our model G and Avin's et al. model A using the parameters (distribution of hyperedges cardinalities, vector M, matrix P) gathered from hypergraph R and presented in Section 5.5. Figure 5.8 compares modularities of G, A, and R. For R the value α equals 0.21. Then the modularity of our model is around 0.69 which is very close to the modularity of R (≈ 0.63). The modularity of A, as A does not feature communities, is very low (≈ 0.06). Figure 5.8 shows also how the modularity of G changes with α and one may notice that it stays at reasonably high level even when the amount of the noise in the network grows.

5.8 Conclusion and Further Work

We have proved theoretically and confirmed experimentally that our model exhibits high modularity, which is rare for known preferential attachment graphs and was not present in hypergraph models so far. While our model has many parameters and may seem complicated, this general formulation allowed us to unify many results known so far. Moreover, it can be easily transformed into much simpler model (e.g., by setting some arguments trivially to 0, repeating the same distributions for hyperedges cardinalities...).

It is commonly known that many real networks present an exponential cut-off in their degree distribution. One possible reason to explain this phenomenon is that nodes eventually become inactive in the network. As further work, we will include this process in our model. The other direction of future study is making the preferential attachment depending not only on the degrees of the vertices but also on their own characteristic (generally called fitness).

Bibliography

[1] C. Avin, Z. Lotker, Y. Nahum, and D. Peleg. Random preferential attachment hypergraph. In Francesca Spezzano, Wei Chen, and Xiaokui Xiao, editors, *ASONAM '19: International Conference on Advances in Social Networks Analysis and Mining, Vancouver, British Columbia, Canada, 27-30 August, 2019*, pages 398–405. ACM, 2019.

[2] A.L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[3] V.D. Blondel, J.L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *J. Stat. Mech. - Theory E*, 2008(10):P10008, 2008.

[4] M. Bloznelis, E. Godehardt, J. Jaworski, V. Kurauskas, and K. Rybarczyk. Recent progress in complex network analysis: Models of random intersection graphs. In Berthold Lausen, Sabine Krolak-Schwerdt, and Matthias Böhmer, editors, *Data Science, Learning by Latent Structures, and Knowledge Discovery*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 69–78. Springer, 2013.

[5] B. Bollobás and O. Riordan. *Handbook of Graphs and Networks: From the Genome to the Internet*. Wiley-VCH, 2003. Pages 1–34.

[6] U. Brandes, D. Delling, M. Gaertler, R. Gorke, M. Hoefer, Z. Nikoloski, and D. Wagner. On modularity clustering. *IEEE Transactions on Knowledge and Data Engineering*, 20(2):172–188, 2008.

[7] P.G. Buckley and D. Osthus. Popularity based random graph models leading to a scale-free degree sequence. *Discrete Math.*, 282(1-3):53–68, 2004.

[8] F. Chung and L. Lu. The average distances in random graphs with given expected degrees. *P. Natl. Acad. Sci. USA*, 99(25):15879–15882, 2002.

[9] F. Chung and L. Lu. *Complex Graphs and Networks*. American Mathematical Society, 2006.

[10] C. Cooper and A.M. Frieze. A general model of web graphs. *Random Struct. Algor.*, 22(3):311–335, 2003.

[11] S. Fortunato. Community detection in graphs. *Phys. Rep.*, 486:75–174, 2010.

[12] S. Fortunato and D. Hric. Community detection in networks: A user guide. *Phys. Rep.*, 659:1–44, 2016.

[13] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *P. Natl. Acad. Sci. USA*, 99(12):7821–7826, 2002.

[14] W. Hoeffding. Probability inequalities for sums of bounded random variables. *J. Am. Stat. Assoc.*, 58(301), 1963.

[15] E. Jacob and P. Mörters. Spatial preferential attachment networks: Power laws and clustering coefficients. *Ann. Appl. Probab.*, 25(2):632–662, 04 2015.

[16] M. Kaiser and C.C. Hilgetagr. Spatial growth of real-world networks. *Phys. Rev. E*, 69:036103, 2004.

[17] B. Kamiński, V. Poulin, P. Prałat, P. Szufel, and F. Théberge. Clustering via hypergraph modularity. *Plos One*, 14:e0224307, Feb 2019.

[18] A. Lancichinetti and S. Fortunato. Limits of modularity maximization in community detection. *Phys. Rev. E*, 84:066122, 2011.

[19] M. Mitzenmacher and E. Upfal. *Probability and Computing: Randomization and Probabilistic Techniques in Algorithms and Data Analysis*. Cambridge University Press, USA, 2nd edition, 2017.

[20] M. Molloy and B.A. Reed. A critical point for random graphs with a given degree sequence. *Random Struct. Algor.*, 6(2/3):161–180, 1995.

[21] N. Neubauer and K. Obermayer. Towards community detection in k-partite k-uniform hypergraphs. In *Proceedings of the NIPS 2009 Workshop on Analyzing Networks and Learning with Graphs*, 2009.

[22] M.E.J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69:026113, Feb 2004.

[23] Przemysław Szufel. Simplehypergraphs. https://github.com/pszufe/SimpleHypergraphs.jl.

[24] V.A. Traag, L. Waltman, and N.J van Eck. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep. UK*, 9(5233), 2019.

[25] J.W. Wang, L.L. Rong, Q.H. Deng, and J.Y. Zhang. Evolving hypernetwork model. *Eur. Phys. J. B*, 77:493–498, 2010.

[26] D. Watts and S. Strogatz. Collective dynamics of small-world networks. *Nature*, 393:440–442, 1998.

[27] W. Yang, G. Wang, Md. Z. A. Bhuiyan, and K.K.R. Choo. Hypergraph partitioning for social networks based on information entropy modularity. *J. Netw. Comput. Appl.*, 86:59–71, 2017.

Chapter 6

Revisiting Preferential Attachment with Applications to Online Social Networks like Twitter

6.1 Introduction

Of chief importance in network algorithmics is the testing of the new algorithms on some representative benchmarks before deploying them at large scale. Very often, such benchmarks are difficult, or even impossible to obtain, that is why we also need to create synthetic ones. Various graph models have been proposed over the years in order to describe real-life complex networks. So far, the networks considered have ranged from physical structures (percolation), to road networks (random geometric graphs), telecommunication networks, online social networks (*e.g.*, Facebook), etc. Most of the networks representing social activities exhibit common properties such as heavy-tailed degree distributions [11], small average distances [5], community structures [20, 23], etc. In this work, we focus on the analysis and the simulation of *preferential attachment models* – for which we will first give a brief state of the art.

6.1.1 Related work

In this Chapter, we only focus on models with linear preferential attachment. For the sake of simplicity, we therefore use the common definition of preferential attachment in this Chapter: roughly, a *preferential attachment model* is defined as a model in which, for a given evolving graph G, the probability for a node in G to form a new edge depends linearly on its degree. The so-called preferential attachment paradigm has been known for about a century [30], but it has started attracting attention since the seminal work of Albert and Barabási [6], in which the authors proposed an explanation to the structure and the degree sequence generated by general human inclinations. Since then, it has shown to be a great tool to describe complex networks. In particular, it has been used since the 2000's in order to describe the evolution of the World-Wide Web [8]. More recently, preferential attachment processes have gained importance in the study of online social networks, where they can be used in order to observe, understand, influence or even manufacture social

phenomena.

Many papers have proposed more sophisticated random models based on the preferential attachment paradigm (see, e.g., [12, 19]). Unfortunately, while in doing so we may better capture the properties of complex networks, the resulting equations for these models become much harder to solve. To take a concrete example, *in the remainder of this article we will exclusively focus on the study of degree sequences in some random (di)graph models.* However, our techniques are quite general, and they could also be applied to the study of other parameters such as the clustering. Let us first digress on *directed* social networks such as Twitter – whose analysis was one of our initial motivations for this work. So far, only a few preferential attachment models in the literature are for *directed* graphs [9]. Still, some social networks are inherently directed, and among them Twitter which is based on *followers* and *followed*. Twitter has only been recently explored [14, 22] and our results show that the existing models failed to reproduce some of its main features, especially the large amount of bidirectional links and the correlation between degree distributions. In order to better explain this peculiar structure, we introduce in this work a new preferential attachment model taking into account these properties. While analysing our new model, some complications have occurred due to the correlations between the out-degree distribution and the distribution of bidirectional links. This led us to look for a way to compute these distributions for more general preferential attachment processes.

Our main contribution is to prove deeper connections between the analysis of preferential attachment models and some associated homogeneous *continuous-time Markov processes*. In many cases of interest, these associated processes are generalized birth-death processes [1], or at least simple enough so that we can derive a closed formula for their stationary distributions. Notably, among many other implications, these new findings allowed us to circumvent the aforementioned difficulties in the study of our Twitter random model.

The relationship between preferential attachment and Markov chains is not that new. In fact, some researchers now present their own variations of preferential attachment directly within the framework of Markov theory [28]. Nevertheless, we found that a rigorous treatment of this relationship was yet to be done. Indeed, starting with the so called "continuum theory" [7], the dominant approach in the literature for showing such a relationship has been mostly empirical, and it could be summarized as follows: first consider a *first-order approximation* of the model (i.e., with some low-order terms removed from the equations); then, find some analogy between these simplified recurrence equations and a differential equation system so as to deduce what the solution should be; finally, validate the results obtained through experiments [25, 24]. We note that this latter approach is a particular case of the differential equation method, that was proved to be correct in many interesting scenarios [29]. Unfortunately, such correctness theorems mostly apply to finite-state random systems or under some Cauchy conditions and they can hardly be generalized, *e.g.*, to the infinite asymptotic degree sequence of a random graph model. In [9], Bollobas et al. astuciously bounded the error term between the degree sequence of their directed preferential attachment model and its first order approximation – thereby making the relationship between the analysis of their model and Markov theory completely rigorous. For that, they took advantage of the fact that, in an ever growing network model (i.e., with no edge-loss), the normalized

degree sequence is given by some infinite *triangular* system. Then, we can upper-bound all the error terms by induction (see also [16] for a similar approach). We note that such a generic method could be applied to many of the classical preferential attachment models from the literature, and even to our Twitter random model – as we did first. However, in practice, its application requires some cumbersome intermediate computations. Furthermore, this approach quickly comes to an end if one wishes to include the possibility of edge removals in these models. To the best of our knowledge, reference [18] is the only rigorous study of a relationship between *some* preferential attachment models with edge deletions and Markov theory. Some of the models considered in this previous work do not fit in our new framework (presented next). However, conversely we have several important models that do not fit in the framework of [18]: such as the undirected model of Chung-Lu [2] and the directed one of Bollobas et al [9]. Indeed, previous works, such as [16, 18], implicitly assume the number of users in the systems (i.e., the order of the random graph) to be deterministic – say exactly equal to t after t discrete time steps. By allowing *edge events* (a.k.a., either a densification or a sparsification of the graph without any modification of its order), this number becomes a random variable and it makes some new low-order error terms appear which we need to consider in our analysis. Our new framework does account for such error terms, thereby making a new step toward a rigorous treatment of all the existing types of preferential attachment random models in the literature.

6.1.2 Contributions

Our key contribution is a *new generic theoretical framework* allowing to compute, somewhat automatically, the first order of the outcome distribution for a large class of random processes that captures and generalizes preferential attachment processes. In particular, all these processes are proved to be stationary up to a rescaling. This allows us to reduce the problem of determining the degree distribution to the much simpler computation of the fixed point of a continuous time Markov process. We then show, both theoretically and by extensive simulations, that, for cases for which a closed-form formula is out of reach, this fixed point can be well estimated using a finite Markov chain acting on truncations of the distribution.

This approach can be applied to many of the existing preferential attachment models, even to some variants with *degree correlations* and *edge removals* that have been so far underexplored in the literature. This novel technique has similarities with the one used by Bollobás et al. in [9] (based on linear operators), and the so-called continuum theory [7]. In order to go beyond these prior analyses, we need to further refine the analysis of low-order terms in our equations and to introduce several intermediate results on the so-called Markov processes with restart [4].

We summarize our work as follows.

- The main contribution is *a new theoretical framework* enabling to study a large class of density distributions, including the most standard models of preferential attachment processes and many new variations. The broad literature on Markov chains can thus be applied to the study of many random graph models.

In order to illustrate the resulting benefits, we propose the following side contributions:

- We first apply our framework to analyze standard models of the field, providing their degree distributions in a fast and unified way.
- We also provide an *improved analysis* of the degree distributions in the classical undirected preferential attachment models and in the directed random model from [9]. For the latter, we believe that we are the first to *determine the joint distribution* of in- and out-degrees (Proposition 6). Until now, only the marginal distributions were known.
- We then use the framework to study two preferential attachment models which are hard to analyze with current methods: one, because of the correlation between its degrees, the other, because edge removals imply that one must use a fixed point equation instead of a direct induction. For both models, we validate the framework by comparing its results with the ones obtained by extensive simulations.
- To introduce an accurate model for Twitter, we first report our *experiments on the degree distributions* in the Twitter graph. Our analysis was performed on the same dataset as for [13] which was made available by the authors. The collected graph has 505 million nodes and 23 billion edges. We observe that a *constant* fraction of the arcs in Twitter are bidirectional, and that their distribution is correlated to the out-degree distribution.
- Lastly, based on our experimental results, we propose a *new preferential attachment model* for directed graphs. We show, through an analysis of its degree distributions, that our model better accounts for the specific properties of Twitter.

Organization. This Chapter is organized as follows. After having introduced the necessary requirements in Sec. 6.2, our new framework for studying the preferential attachment processes is formally presented in Section 6.3. We then discuss in Section 6.4 how the stationary distribution of the introduced infinite Markov process can be efficiently computed by considering finite-state processes obtained by truncation. In Section 6.5, we then apply our framework to derive the degree distributions of classic models, first, with a one-dimensional state space and then, with a two-dimensional state space. In Section 6.6, we present the experiments that led us to discover peculiar empirical properties of Twitter. Then, we describe and motivate our new random model, and provide its joint degree distribution. We also validate the framework through simulations. We conclude in Section 6.7.

6.2 Preliminaries

In this section, we focus on a large class of density distributions that are properly defined in Sec. 6.2.1. Roughly, this class is a far-reaching generalization of the degree distributions for Preferential Attachment (PA) random graph models. We show the link between the latter models and our class of density distributions through a series of examples.

Finally for all the distributions considered, we introduce an associated continuous-time Markov process in Sec 6.2.2, and some intermediate results on Markov theory.

6.2.1 Assumptions on the distribution

Let $\mathcal{S}$ be a countable space. A *density distribution* on $\mathcal{S}$ is a discrete-time process, defined at any time step $t \in \mathbb{N}_{\geq t_0}$ as a sequence $x(t) := (x_i(t))_{i\in\mathcal{S}}$ of nonnegative random variables indexed by $\mathcal{S}$. Such a process defines a population of size $s(t) := \sum_{i\in\mathcal{S}} x_i(t)$ individuals. For every $i \in \mathcal{S}$ we say that there are $x_i(t)$ individuals in state i*. In the particular case of preferential attachment models, this state represents the degrees of the vertex. As examples, we take $\mathcal{S} = \mathbb{N}$ for undirected graphs, whereas we take $\mathcal{S} = \mathbb{N}^2$ for digraphs — because each vertex now has an in-degree and an out-degree. Choosing $\mathcal{S} = \mathbb{N}^k$, for some $k > 2$, may help in modelling more complex situations. For instance, in order to model Twitter (Sec. 6.6), we will take $\mathcal{S} = \mathbb{N}^3$ in order to better account for bidirectional edges.

In the remainder of this Chapter, we only consider the density distributions on $\mathcal{S}$ that obey the following additional properties:

Property 1. *$s(t)$ follows a* mean concentrated distribution *such as e.g., a Bernoulli distribution or a Poisson point process. Furthermore at any time $t \geq t_0$, the expected size $\mathbb{E}[s(t)]$ of the population is exactly t.*

Note that we could assume more generally that $\mathbb{E}[s(t)] = c \cdot t$, for some positive c. Actually, as this will be shown by our examples (Sec. 6.2.1), we have for most PA models $\mathbb{E}[s(t)] = c \cdot t$ for some $c < 1$. However, we can always enforce $c = 1$ up to a "rescaling" (i.e., by considering the process $(x_i(t)/c)_{i\in\mathcal{S}}$).

Property 2. *When a new individual appears, its initial state is chosen according to some distribution $\mu : \mathcal{S} \to (0; 1)$. We assume that μ has finite support, or equivalently the number of entering states for the individuals is finite.*

In the particular case of PA processes, the addition of a new vertex is called a "node event". Two node events may differ in the way connections are made between the new node and the existing ones. If there are only a finite number of node events, then the number of possible entering states for an individual is finite, and so Property 2 is always true.

Property 3. *There is some fixed universal constant K such that, at time $t \geq t_0$, there are no more than K individuals leaving their state for another.*

Recall that for PA models, the state of a node represents its degree(s). Then, this above Property 3 follows from the fact that only a *constant-number* of new edges are added or removed at each step.

Property 4. *At time $t \geq t_0$, an individual in state i has some probability $p_{i,j}(t)$ to change to state j. We assume that all these transitions probabilities are determined by some "evolution function" $e(t)$: with the latter being just a tuple of random variables. Namely, we assume the existence of some fixed sequence $(m_{i,j})_{i,j\in\mathcal{S},i\neq j}$, and of some function f, such that the following hold:*

*We may think of the $x_i(t)$'s as integer variables. However, some of the operations that we use, e.g., scaling, may transform them into some nonnegative real numbers.

- $p_{i,j}(t) = f(i, j, e(t))$ – *the transition solely depends on* $e(t)$ *and on the states* i *and* j*;*
- *and* $p_{i,j}(t)$ *is mean-concentrated, with its mean being equal to* $\frac{m_{i,j}}{t}$*. More formally, let* $\upsilon_{i,j}(t) = c \cdot \left(\frac{m_{i,j}}{t}\right)^{1+d}$*, for some universal constants* c *and* d*; we have:*

$$\mathbb{P}\left[\ \left|p_{i,j}(t) - \frac{m_{i,j}}{t}\right| \leq \upsilon_{i,j}(t)\right] \geq 1 - t^{-d}.$$

For example, in the case of preferential attachment processes, we may think of the function e as a record of all types of node and edge events that have happened until then – or more simply, as a pair that contains the number of nodes and the number of edges. Indeed, changes of state correspond to edge addition or removal. Then, we have (up to lower-order terms) $p_{i,j}(t) = m_{i,j}/N(t)$, where the function $N(t)$ is a (normalized) linear function of the number of vertices and of the number of edges. We use the error function υ in order to take account of: `(i)` the mean deviation of $N(t)$ (that depends on the number of each different events that occured, e.g., how many node events, edge additions or deletions); and `(ii)` the small probability to select a same vertex twice for an edge event, *i.e.*, to create loops, typically of order $\mathcal{O}(t^{-2})$.

Property 5. The transitions are local. *For instance if* $\mathcal{S} = \mathbb{N}^k$*, then for any* i, j *s.t.* $m_{i,j} > 0$*, we may assume that the (Manhattan) distance between* i *and* j *is bounded. More generally, we assume the existence of a sequence of increasing finite subsets* $(B_n)_{n\in\mathbb{N}^*}$ *and of a constant* c_B *s.t.:*

1. $\bigcup_{n\geq 1} B_n = \mathcal{S}$*;*
2. *and for every* $n \geq 1$ *and for every* $i \in B_n$*, we have* $m_{i,j} \neq 0 \Longrightarrow j \in B_{n+c_B} \setminus B_{n-(c_B+1)}$*.*

Observe that we can assume w.l.o.g. $c_B = 1$*, and that the support of* μ *is contained in* B_1*.*

Property 5 has a natural interpretation for PA models. Specifically, it also follows from the fact that at any time step t, we can only add and/or remove a constant number of edges. Then, a natural choice for B_n would be the set of vertices of total degree at most n.

We stress that the behaviour of our processes mostly depend (up to lower-order terms) on $\mathcal{S}, \mu$ and m. Therefore, we will abusively refer to such a process as the triple $(\mathcal{S}, m, \mu)$ in what follows.

Examples

• Consider first the classical Barabasi-Albert model [6]. At every time step, a new node is added and made adjacent to M existing nodes that are chosen with probability linearly proportional to their degree. Note that, for any time step t, we have $s(t) = t$ and so we do not need to rescale in this case. Furthermore the sum of the degrees is equal to $2Mt$. Therefore for every $j \in \{1, \ldots, M\}$ we have the following transition probability:

$$p_{i,i+j}(t) = \binom{M}{j} \left(\frac{i}{2Mt}\right)^j \left(1 - \frac{i}{2Mt}\right)^{M-j}.$$

We observe that if $M > 1$ and $j > 1$ then the latter probability is just a low-order term that is caused by double events. It is roughly upper-bounded by $2^M \left(\frac{i}{2Mt}\right)^2$. Otherwise we have $p_{i,i+1}(t) = i/(2t) + \mathcal{O}((i/(2t))^2)$. Overall, we obtain a scale-invariant process $(\mathbb{N}, m, \mu)$ such that: $\mu_M = 1$, and for every $i, j \in \mathbb{N}$ we have that:

$$\begin{cases} m_{i,i+1} = i/2 \\ m_{i,j} = 0 \text{ if } j \neq i+1. \end{cases}$$

- Consider now the random model introduced by Chung and Lu in [10]. At every time step, with some probability p, a new node is added and made adjacent to an existing node that is chosen with probability linearly proportional to its degree. Otherwise, with probability $1-p$, two existing nodes are chosen with probability linearly proportional to their degree and made adjacent.

Here, for any time step t, the sum of the degrees is equal to $2t$, while the number of nodes is concentrated around pt. So, in order to force $\mathbb{E}[s(t)] = 1$ we need to rescale. We get the scale-invariant process $(\mathbb{N}, m, \mu)$ such that $\mu_1 = 1$, and for every $i, j \in \mathbb{N}$ we have that:

$$\begin{cases} m_{i,i+1} = pi/2 + (1-p) \cdot 2 \cdot i/2 = (1 - p/2)\, i \\ m_{i,j} = 0 \text{ if } j \neq i+1. \end{cases}$$

- The random model of Bollobás et al. [9] is a bit more complex to analyse. Let $\alpha, \beta, \gamma, \delta_{in}, \delta_{out}$ be positive constants such that $\alpha + \beta + \gamma = 1$. We denote by $d^-(v)$ and $d^+(v)$, respectively, the in-degree and the out-degree of a given node v. For every time step t, we have that: with probability α, a new vertex v is created with an arc from v to an existing vertex w – chosen according to $d^-(w) + \delta_{in}$; with probability γ, a new vertex v is created with an arc from an existing vertex w – chosen according to $d^+(w) + \delta_{out}$ – to v; otherwise, with probability β, a new arc is added from an existing vertex u – chosen according to $d^-(u) + \delta_{in}$ – to an existing vertex w – chosen according to $d^+(w) + \delta_{out}$.

Note that, for any time t, the number of arcs is exactly t. Furthermore, by Chernoff bounds, we have that the number of nodes $s(t)$ is concentrated around its mean $(1-\beta) \cdot t$. So, after rescaling the node events (division by $\alpha + \gamma = 1 - \beta$), one finally obtains the following scale-invariant process $(\mathbb{N}^2, m, \mu)$ such that: $\mu(0,1) = 1 - \mu(1,0) = \alpha/(1-\beta)$, and for every $(i,j) \in \mathbb{N}^2$ we have that:

$$\begin{cases} m_{(i,j),(i+1,j)} = \frac{(\alpha+\beta)(i+\delta_{in})}{1+(1-\beta)\cdot\delta_{in}} \\ m_{(i,j),(i,j+1)} = \frac{(\gamma+\beta)(j+\delta_{out})}{1+(1-\beta)\cdot\delta_{out}} \\ m_{(i,j),(k,l)} = 0 \text{ otherwise.} \end{cases}$$

6.2.2 Markov Process with a Reset

We refer to [15] for a standard reference book on continuous-time Markov processes. For any conservative q-matrix A (*a.k.a.*, transition rate matrix), we denote by $X_A(t)$ the corresponding homogeneous Markov process. Let $P_A(t) = (p_A(i,j,t))_{i,j \in \mathcal{S}}$, where $p_A(i,j,t) = \mathbb{P}r[X_A(t) = j \mid X_A(0) = i]$. This parametric family of matrices (indexed by t) satisfies both the forward and backward Kolmogorov equations, namely:

$$P'_A(t) = P_A(t)A = AP_A(t).$$

Given some initial distribution μ, let us also define $p_A(t) = (p_A(i,t))_{i\in\mathcal{S}}$ as the state distribution of the process X_A, where $p_A(i,t) = \mathbb{P}r[X_A(t) = i \mid X_A(0) \sim \mu] = \sum_r \mu_r \cdot p_A(r,i,t)$.

The relationship between PA models and Markov theory is as follows. To any distribution $(\mathcal{S}, m, \mu)$ as earlier defined in Sec. 6.2.1, we can associate to it the following continuous-time Markov process:

Definition 5. *For $(\mathcal{S}, m, \mu)$, the associated (continuous-time) Markov process on the state space $\mathcal{S}$ has transition rate matrix $Q = (q_{i,j})_{i,j\in\mathcal{S}}$ such that:*

- *for every $i \neq j$, $q_{i,j} = m_{i,j} + \mu_j$;*
- *for every i, $q_{i,i} = -\sum_{j\neq i} q_{i,j}$ (every line sums to zero).*

At first glance, it might be not so clear why this above sum $\sum_{j\neq i} q_{i,j}$ should converge. However, for every fixed i we can deduce from Property 5 that there can be only a *finite* number of nonzero terms $m_{i,j}$. Furthermore, there is a finite number of nonzero μ_j (Property 2). Overall, the number of nonzero $q_{i,j}$ is finite, and so, $q_{i,i}$ is well-defined.

The remainder of this section is now devoted to some intermediate results on Markov processes, and especially those with a reset [4] of which the associated process given by Def. 5 is a particular case.

Non-explosiveness.

The process $\{X_A(t) \mid t \in (0;\infty)\}$ is called *non-explosive* if within any finite time t, the number of jumps is finite with probability 1. A necessary and sufficient condition for being non-explosive is that, with probability 1, the average time spent in the n first states goes to infinity as n grows [27]. There is another definition for this concept in terms of a moment drift function. Specifically, for a conservative q-matrix A over the countable state space $\mathcal{S}$, the function $f_A : \mathcal{S} \to (0;\infty)$ is a *drift function* if there exists some constants D_1, D_2 such that we have, for any $i \in \mathcal{S}$:

$$\sum_{j\in S} a_{i,j} \cdot f_A(j) \leq D_1 \cdot f_A(i) \text{ and } -a_{i,i} = \sum_{j\neq i} a_{i,j} \leq D_2 \cdot f_A(i).$$

It is a *moment function* if we also have, for any increasing sequence of finite subsets $(B_n)_{n\in\mathbb{N}}$ such that $\bigcup_{n\in\mathbb{N}} B_n = \mathcal{S}$, $\lim_{n\to\infty} \inf_{i\notin B_n} f_A(i) = +\infty$. A process is non-explosive if and only if it has a drift moment function [26].

Invariant measure.

Next, given some conservative q-matrix M and some initial distribution μ, we denote by Q the transition-rate matrix of the same process as defined by M but with reset to the initial distribution μ at rate 1.

We will need the following results:

Lemma 19 ([4])**.** $p_Q(i,j,t) = e^{-t} \cdot p_M(i,j,t) + \int_0^t \sum_{r\in\mathcal{S}} e^{-s} \cdot \mu_r \cdot p_M(r,j,s)ds$.

Lemma 20 ([4]). *The distribution π s.t., for every $i \in \mathcal{S}$ we have $\pi_i = \sum_r \mu_r \int_0^\infty e^{-t} \cdot p_M(r,i,t)dt$ is the unique invariant measure for Q. Moreover, for any $i,j \in \mathcal{S}$ we have:*

$$|p_Q(i,j,t) - \pi_j| \leq e^{-t}.$$

Note that Lemmas 19 and 20 are proved in [4] under the assumption that M is honest, that is a weaker property than being non-explosive [15].

Corollary 3. $p_Q(t) = e^{-t} \cdot p_M(t) + \int_0^t e^{-s} \cdot p_M(s)ds$.

Proof. For every $i \in \mathcal{S}$, we have $p_Q(i,t) = \sum_r \mu_r \cdot p_Q(r,i,t)$. By Lemma 19:

$$\begin{aligned}
&\sum_r \mu_r \cdot p_Q(r,i,t) \\
&= \sum_r \mu_r \cdot \left[e^{-t} \cdot p_M(r,i,t) + \int_0^t \sum_{r' \in \mathcal{S}} e^{-s} \cdot \mu_{r'} \cdot p_M(r',i,s)ds \right] \\
&= \left[e^{-t} \cdot \sum_r \mu_r \cdot p_M(r,i,t) \right] + \sum_r \mu_r \cdot \int_0^t e^{-s} \cdot \sum_{r'} \mu_{r'} \cdot p_M(r',i,s)ds \\
&= e^{-t} \cdot p_M(i,t) + \sum_r \mu_r \cdot \int_0^t e^{-s} \cdot p_M(i,s)ds \\
&= e^{-t} \cdot p_M(i,t) + \int_0^t e^{-s} \cdot p_M(i,s)ds.
\end{aligned}$$

□

6.3 Framework on Preferential attachment processes

Given a process $(\mathcal{S}, m, \mu)$ as in Sec. 6.2.1, let us consider the normalized state distribution $\bar{x}_i(t) = x_i(t)/s(t)$ where $x_i(t)$ is the number of individuals in state i at time $t \geq 0$ and $s(t)$ is the total number of individuals. In what follows, we aim at computing the first order of this distribution. We prove, in Sec. 6.3.1, that under some technical assumptions (always satisfied for PA models), the latter is equal to the invariant measure of an associated Markov process (given by Def. 5). Note that in the literature, computing the degree distributions of random graph models usually requires the technical use of concentration inequalities (in an ad-hoc way depending on the model), and the tedious solving of a recurrence equation [9]. Our generic framework (Theorem 7) improves upon both aspects. Specifically, we completely automate the use of concentration inequalities for a broad range of scale-invariant processes. Furthermore, in some cases where the associated process is well-understood, our reduction considerably simplifies the solving of our recurrence equations.

6.3.1 Reduction to a Markov process

For a given process $(\mathcal{S}, m, \mu)$ let Q be as in Def. 5. This above Markov process admits a unique invariant measure π that follows from previous works on Markov processes with reset [4] — see Sec. 6.2.2 for details. In the remainder of this section,

we prove that under some technical assumptions on the coefficients $m_{i,j}$ (always satisfied for PA models), there is *convergence in law* of $\bar{x}(t)$ to π, i.e. $\lim_{t\to\infty} \bar{x}_i(t) = \pi_i$ for every state i.

Theorem 7. *If the following conditions are true then there is convergence in law of $\bar{x}(t)$ to the unique invariant measure for Q:*

1. *There exists a $n_0 \geq 1$ s.t., for every $n \geq n_0$ and $i \in B_n$, $\sum_{j\neq i} m_{i,j} \leq n-1$;*
2. *For every $n \geq 1$ and $i \in B_n \setminus B_{n-1}$, $\sum_{j\notin B_n} m_{j,i} \leq \sum_{j\in B_{n-1}} m_{j,i}$.*
3. *For some choice of positive λ we have, for every $i \in \mathcal{S}$: $\sum_{j\neq i} m_{j,i} \leq \left(\sum_{j\neq i} m_{i,j}\right) - 2\lambda$.*

We stress that all these above conditions have a very intuitive interpretation for PA models. Indeed, we say with Cond. 1 that an individual changes her state at a rate which is slightly sub-linear in her degree. Very roughly, Cond. 2 ensures that it is more likely for an individual to increase her degree rather than to decrease it. Cond. 3 implies that we leave any state i at a bigger rate than the one at which we are entering in i — in other words, we favour the vertices with large degree. However, in general, we see no reason why these conditions would be necessary for having convergence in law to the invariant measure of Q. We leave as an interesting open question what could be the necessary and sufficient conditions for having such a convergence.

The rest of this Section is dedicated to prove Theorem 7. It is organized as follows:

- First, we establish a recurrence relation for $(\mathbb{E}[\bar{x}(n)])_n$ in Sec. 6.3.2 that involves the transition rate matrix Q.
- Then, we explain why we can restrict our analysis to the normalized *mean state distribution* $\mathbb{E}[\bar{x}_i(t)]$ (Sec. 6.3.3). For that, we use classical concentration inequalities.
- We continue by proving intermediate relationships between the processes defined by Q and $M = (m_{i,j})_{i,j\in\mathcal{S}}$, respectively, where $m_{i,i} \underset{def}{=} -\sum_{j\neq i} m_{i,j}$ (Sec. 6.3.4). Here the key observation is that M defines the same process as Q, but conditioned on the fact that there is no reset to the initial distribution μ. For the process defined by Q, this reset happens at rate 1.
- Finally, in Sec. 6.3.5, we bound the absolute difference between $(\mathbb{E}[\bar{x}(n)])_n$ – at *discrete time* step n – and the state distribution of the process defined by Q at some *continuous time* $T_n = \Theta(\log n)$ (exponentially smaller than n).

6.3.2 The Recurrence Equation

We start by making explicit the relationship between $\mathbb{E}[\bar{x}(n)]$ and the matrix Q.

Lemma 21. *For some positive constants c', d and for every $i \in \mathcal{S}$:*

$$\left| \mathbb{E}[\bar{x}_i(n+1)] - \left(\mathbb{E}[\bar{x}_i(n)] + \frac{1}{n+1} \cdot (\mathbb{E}[\bar{x}(n)]Q)_i \right) \right| \leq c' \cdot \left(\frac{\sum_{j\neq i} m_{i,j} + m_{j,i}}{n} \right)^{1+d}.$$

Proof. By the hypothesis (*i.e.*, Sec. 6.2.1):

$$\begin{aligned}\mathbb{E}[x_i(n+1) \mid x(n), e(n)] = x_i(n) &+ \sum_{j \neq i} p_{j,i}(n) \cdot x_j(n) \\ &- \left[\sum_{j \neq i} p_{i,j}(n)\right] \cdot x_i(n) \\ &+ \mu_i.\end{aligned} \tag{6.1}$$

Recall that we have $\mathbb{E}[\mathbb{E}[x_i(n+1) \mid x(n), e(n)]] = \mathbb{E}[x_i(n+1)]$. We so obtain by taking the expectation:

$$\begin{aligned}\mathbb{E}[x_i(n+1)] = \mathbb{E}[x_i(n)] &+ \sum_{j \neq i} \mathbb{E}[p_{j,i}(n) \cdot x_j(n)] \\ &- \left[\sum_{j \neq i} \mathbb{E}[p_{i,j}(n) \cdot x_i(n)]\right] \\ &+ \mu_i.\end{aligned} \tag{6.2}$$

In order to simplify this above equation, let us first consider $\mathbb{E}[p_{j,i}(n) \cdot x_j(n)]$. We use a trick from [9]. By the Property 3 we have $p_{j,i}(n) \cdot x_j(n) \leq K = \mathcal{O}(1)$. Since $p_{j,i}(n)$ is mean-concentrated (Property 4) then we get that:

$$\mathbb{E}[p_{j,i}(n) \cdot x_j(n)] = \left(\frac{m_{j,i}}{n} + \mathcal{O}\left(\left(\frac{m_{j,i}}{n}\right)^{1+d}\right)\right) \cdot \mathbb{E}[x_i(n)] + \mathcal{O}(n^{-d}). \tag{6.3}$$

This can be simplified using that $\mathbb{E}[x_i(n)] \leq \mathbb{E}[s(n)] = n$ (Property 1). We obtain $\mathbb{E}[p_{j,i}(n) \cdot x_j(n)] = \frac{m_{j,i}}{n} \cdot \mathbb{E}[x_i(n)] + \mathcal{O}\left(\frac{m_{j,i}^{1+d}}{n^d}\right)$. Overall the equation can now be written as:

$$\begin{aligned}\mathbb{E}[x_i(n+1)] = \mathbb{E}[x_i(n)] &+ \sum_{j \neq i} \frac{m_{j,i}}{n} \cdot \mathbb{E}[x_j(n)] \\ &- \left[\sum_{j \neq i} \frac{m_{i,j}}{n}\right] \cdot \mathbb{E}[x_i(n)] \\ &+ \mu_i \\ &+ \varepsilon_i(n),\end{aligned} \tag{6.4}$$

where $\varepsilon_i(n)$ is an error term s.t. $|\varepsilon_i(n)| = \mathcal{O}(\frac{\sum_j (m_{i,j}+m_{j,i})^{1+d}}{n^d})$.

We now normalize the equation by using the same trick as above. By Property 1, $s(n)$ is mean-concentrated and $\mathbb{E}[s(n)] = n$. Moreover, $\bar{x}_i(n) \leq 1 = \mathcal{O}(1)$ in any case. Therefore, we have:

$$|\mathbb{E}[\bar{x}_i(n)] - \mathbb{E}[x_i(n)]/n| = \mathbb{E}[x_n \cdot |1/s(n) - 1/n|] = \mathcal{O}(n^{-d}). \tag{6.5}$$

We divide all terms in the equation by n, that leads to:

$$(n+1) \cdot \Delta_i(n) = -\mathbb{E}[\bar{x}_i(n)] + \sum_j m_{j,i} \cdot \mathbb{E}[\bar{x}_j(n)] + \mu_i + \eta_i(n), \tag{6.6}$$

where $\Delta_i(n) = \mathbb{E}[(\bar{x}_i(n+1) - \bar{x}_i(n))]$, and the error term $\eta_i(n)$ satisfies:

$$|\eta_i(n)| = \mathcal{O}\left(\frac{\left(\sum_{j\neq i} m_{j,i}\right) + \left(\sum_{j\neq i} m_{i,j}\right)}{n^d}\right) + \mathcal{O}\left(\frac{\left(\sum_{j\neq i} m_{i,j} + m_{j,i}\right)^{1+d}}{n^d}\right)$$
$$\leq c' \cdot \frac{\left(\sum_{j\neq i} m_{i,j} + m_{j,i}\right)^{1+d}}{n^d}, \tag{6.7}$$

for some universal constant c'. Then, by using the fact that $\sum_i \mathbb{E}[\bar{x}_i(n)] = 1$, the above equation can be rewritten as follows:

$$(n+1)\cdot\Delta_i(n) = \sum_j m_{j,i}\cdot\mathbb{E}[\bar{x}_j(n)] - \mathbb{E}[\bar{x}_i(n)] + \mu_i \tag{6.8}$$
$$+ (\mu_i\cdot\mathbb{E}[\bar{x}_i(n)] - \mu_i\cdot\mathbb{E}[\bar{x}_i(n)]) + \eta_i(n)$$
$$= \sum_j m_{j,i}\cdot\mathbb{E}[\bar{x}_j(n)] - (1-\mu_i)\cdot\mathbb{E}[\bar{x}_i(n)] \tag{6.9}$$
$$+ (1-\mathbb{E}[\bar{x}_i(n)])\cdot\mu_i + \eta_i(n)$$
$$= \sum_j m_{j,i}\cdot\mathbb{E}[\bar{x}_j(n)] - \left(\sum_{j\neq i}\mu_j\right)\cdot\mathbb{E}[\bar{x}_i(n)] \tag{6.10}$$
$$+ \sum_{j\neq i}\mu_i\cdot\mathbb{E}[\bar{x}_j(t)] + \eta_i(n)$$
$$= \sum_{j\neq i}(m_{j,i}+\mu_i)\cdot\mathbb{E}[\bar{x}_j(n)] \tag{6.11}$$
$$- \left(\sum_{j\neq i} m_{i,j} + \mu_j\right)\cdot\mathbb{E}[\bar{x}_i(n)] + \eta_i(n)$$
$$= \sum_j q_{j,i}\cdot\mathbb{E}[\bar{x}_j(n)] + \eta_i(n) \tag{6.12}$$
$$= (\mathbb{E}[\bar{x}(n)]Q)_i + \eta_i(n). \tag{6.13}$$

Overall we obtain:

$$\left|\ \Delta_i(n) - \frac{1}{n+1}\cdot(\mathbb{E}[\bar{x}(n)]Q)_i\ \right| \leq \frac{|\eta_i(n)|}{n+1} \tag{6.14}$$
$$\leq c'\cdot\left(\frac{\sum_{j\neq i} m_{i,j} + m_{j,i}}{n}\right)^{1+d}. \tag{6.15}$$

□

6.3.3 Mean Concentration

Lemma 22. *There is convergence in law of $x(n)$ to $\mathbb{E}[x(n)]$.*

Proof. Our approach closely follows the one from [9]. First let $i \in \mathcal{S}$ be fixed, and let $\mathcal{E}$ be a random outcome for the infinite sequence $(e(n))_n$. Since $s(n)$ (Property 1) and the transition probabilities (Property 4) are mean-concentrated, the sequence

$(\mathbb{E}[\bar{x}(n) \mid \mathcal{E}])_n$ satisfies the recurrence equation of Lemma 21 with probability 1 — up to some changes of constant for the lower-order terms. Therefore, we are left to prove the lemma conditioned on $\mathcal{E}$. For that, let $n \geq 0$ be arbitrary. For every $t \in \{0, 1, \ldots, n\}$, we define $X_t = \mathbb{E}[x_i(n) \mid x(t-1), \mathcal{E}]$. Then, $(X_t)_t$ is a martingale, sometimes called a Doob martingale. In particular, if for some constant C we have $|X_{t+1} - X_t| \leq C$, then by Azuma's inequality, we have that $x_i(n) \mid \mathcal{E}$ (*i.e.*, conditioned on $\mathcal{E}$) is concentrated around its mean with high probability [21]. Therefore, in order to prove the lemma, it suffices to prove the existence of such a constant C. We claim that it is a consequence of Property 3. Indeed, between the time-steps t and $t+1$, there is a subset V_t of at most K individuals changing their state. For any other individual (possibly not existing yet at time t), the probability to be in state i at time n only depends on the fixed sequence $\mathcal{E}$ (by Property 4), and in particular, it is independent of V_t. As a result, we have for any v_t^1, v_t^2 that:

$$\begin{aligned} &|\mathbb{E}[x_i(n) \mid x(t-1), V_t = v_t^1, \mathcal{E}] - \mathbb{E}[x_i(n) \mid x(t-1), V_t = v_t^2, \mathcal{E}]| \\ &\leq |v_t^1| + |v_t^2| \leq 2K. \end{aligned} \tag{6.16}$$

This implies that $|X_{t+1} - X_t| \leq 2K$. □

From now on, we are left to study the (normalized) mean state distribution.

6.3.4 Euler method

We now use the results in Sec. 6.2.2 in order to derive some useful properties of the process defined by Q (the process without a reset, defined by M, resp.). In particular, we first observe that these two processes are non-explosive. Indeed, it follows from the Property 5 of locality that the n first steps visited must be in the ball B_n. Furthermore, for every $n \geq n_0$, the holding time for any state in B_n is $\Omega(1/n)$. Therefore, the average time spent in the n first steps is $\Omega(H_n)$, where H_n is the harmonic series.

In what follows, we give an alternative proof in terms of moment drift function.

Lemma 23. *For every positive α, β, the function $f : \mathcal{S} \to (0; +\infty)$ s.t. for every $n \geq 1$ and $i \in B_n \setminus B_{n-1}$, $f(i) = \alpha \cdot n + \beta$, is a drift moment function for M. In particular, the process defined by M is non-explosive.*

Proof. For every $n \geq 1$ and $i \in B_n \setminus B_{n-1}$, we have by the Property 5 of locality that $m_{i,j} \neq 0 \Longrightarrow j \in B_{n+1}$. In particular, we have by Condition 1 of the theorem:

$$\begin{aligned} \sum_j m_{i,j} \cdot f(j) &\leq \left(\sum_{j \neq i} m_{j,i} \right) (\alpha \cdot (n+1) + \beta) + m_{i,i} \cdot (\alpha \cdot n + \beta) \\ &= \alpha \cdot |m_{i,i}| = \mathcal{O}(n). \end{aligned}$$

□

Altogether combined, these above technicalities allow us to apply some Euler method on $p_Q(t)$. Indeed, our whole approach is inspired from the non-rigorous continuum theory [7], and more generally from the differential equation method [29].

Lemma 24. *Let $(T_n)_n$ be an increasing sequence of positive real numbers such that $T_{n+1} - T_n \sim_{+\infty} \frac{1}{n+1}$. For every $i \in \mathcal{S}$ we have:*

$$\left| p_Q(i, T_{n+1}) - \left[p_Q(i, T_n) + \frac{1}{n+1} \cdot (p_Q(T_n)Q)_i \right] \right| \leq o\left(\frac{e^{-T_n}}{n+1} \right).$$

Proof. From the Kolmogorov's Forward Equations, $p'_Q(t) = p_Q(t)Q$. Furthermore for every $i \in \mathcal{S}$ we have by Taylor-Lagrange inequality:

$$|p_Q(i, t+h) - p_Q(i, t) - hp'_Q(i, t)| \leq \frac{h^2}{2} \sup_{\theta \in (0;1)} |p''_Q(i, t + \theta h)|. \tag{6.17}$$

We set $t = T_n$ and $h = h_n = T_{n+1} - T_n$. By the hypothesis, $h_n = (1 + o(1)) \cdot \frac{1}{n+1}$. So, we are left with bounding the following error term:

$$\varepsilon_i(n) = o(n^{-1}) \cdot |p'_Q(i, T_n)| + \frac{(1 + o(1))^2}{2(n+1)^2} \cdot \sup_{\theta \in (0;1)} |p''_Q(i, T_n + \theta h_n)| \tag{6.18}$$

$$= o(n^{-1}) \cdot \left(|p'_Q(i, T_n)| + \sup_{\theta \in (0;1)} |p''_Q(i, T_n + \theta h_n)| \right). \tag{6.19}$$

In order to bound this error term, we use Lemma 20. Specifically:

- For every $i, j \in \mathcal{S}$, we have:

$$|p'_Q(i, j, t)| = |\sum_k q_{i,k} \cdot p_Q(k, j, t)| = |\sum_k q_{i,k} \cdot (p_Q(k, j, t) - \pi_j)| \tag{6.20}$$

$$\leq \sum_k |q_{i,k}| \cdot |p_Q(k, j, t) - \pi_j| \leq \left(\sum_k |q_{i,k}| \right) \cdot e^{-t} \tag{6.21}$$

$$= 2 \cdot |q_{i,i}| \cdot e^{-t} = \mathcal{O}(f(i) \cdot e^{-t}), \tag{6.22}$$

  where f is, for some choices of α and β, the drift moment function for the process that is given by Lemma 23. We stress that $\sum_r \mu_r f(r)$ is a constant because we assume the support of μ to be finite. In particular, for some constant C_1 we have:

$$|p'_Q(i, t)| = |\sum_r \mu_r \cdot p'_Q(r, i, t)| \leq \sum_r \mu_r |p'_Q(r, i, t)| \leq C_1 \cdot e^{-t}. \tag{6.23}$$

- In the same way, for every $i, j \in \mathcal{S}$, we have:

$$|p''_Q(i, j, t)| = |\sum_k q_{i,k} \cdot p'_Q(k, j, t)| \tag{6.24}$$

$$\leq \sum_k |q_{i,k}| \cdot |p'_Q(k, j, t)| \tag{6.25}$$

$$\leq \sum_k |q_{i,k}| \cdot \mathcal{O}(f(k)e^{-t}). \tag{6.26}$$

  In particular, by the Property 5 of locality, we have $q_{i,k} \neq 0 \implies f(k) \leq f(i) + \alpha$, for some positive α. As a result: $|p''_Q(i, j, t)| \leq (\sum_k |q_{i,k}|) \cdot \mathcal{O}(f(i)e^{-t}) = \mathcal{O}(|q_{i,i}| f(i)e^{-t}) = \mathcal{O}(f(i)^2 \cdot e^{-t})$. We stress that $\sum_r \mu_r f(r)^2$ is also a constant since the support of μ is finite. Therefore, we obtain that for some constant C_2: $|p''_Q(i, t)| \leq C_2 \cdot e^{-t}$.

Altogether combined, we get $|\varepsilon_i(n)| \leq o(n^{-1}) \cdot (C_1 + C_2) \cdot e^{-T_n}$. □

6.3.5 Proof of Theorem 7

Proof of Theorem 7
For any undefined notation in what follows, see Sec. 6.2.2. Let $n_1 \geq n_0$ be a large enough integer so that some asymptotic error terms can be neglected in our calculations. For every $n \geq 1$ and $i \in B_n \setminus B_{n-1}$, let $K(i) = C \cdot (n+\beta)^{1+d}$ for some large enough constants C, β (to be fixed by the proof). In what follows, we prove by induction that, for every $n \geq n_1$ and $i \in B_n$ we have:

$$|\mathbb{E}[\bar{x}_n(i)] - p_Q(i, \log(n+1))| \leq \frac{K(i)}{n+1} \cdot \sum_{t=1}^{n-1} \frac{1}{t^d}, \tag{6.27}$$

where d is one of the two absolute constants given by Property 4. We stress that we have:

$$\sum_{t=1}^{n} \frac{1}{t^d} \sim \begin{cases} \zeta(d) = \Theta(1) \text{ if } d > 1 \\ \log n \text{ if } d = 1 \\ \frac{n^{1-d}}{1-d} \text{ otherwise.} \end{cases} \tag{6.28}$$

In particular, we always have $\lim_{n\to\infty} \frac{1}{n+1} \cdot \sum_{t=1}^{n-1} \frac{1}{t^d} = 0$. Therefore, proving the induction hypothesis will prove the theorem. Note that for the base case, B_{n_1} is finite and so for any fixed choice of β, we can always choose a constant C large enough so that this above inequality is valid. From now on, let us assume that this inequality holds for some $n \geq n_1$. We will prove that it also holds for $n+1$.

Claim 7.1.

$$p_Q(B_n, \log(n+1)) \geq 1 - \frac{2e^{-1}}{n+1}.$$

Proof. We first need to lower bound $p_M(B_n, \log(n+1))$ (*i.e.*, the mass in the ball conditioned on the fact that there is no restart). For that, for any $p \geq 1$ let $L_p = B_p \setminus B_{p-1}$. Let $A = (a_{p,q})_{p,q\in\mathbb{N}^*}$ be such that:

$$\begin{cases} a_{p,p+1} = \max_{i\in L_p} \sum_{j\in L_{p+1}} m_{i,j} \\ a_{p,p} = -a_{p,p+1} \\ a_{p,q} = 0 \text{ otherwise.} \end{cases}$$

By construction, $p_A([1;n], t) \leq p_M(B_n, t)$. Therefore, we are left lower-bounding $p_A([1;n], t)$, or equivalently upper-bounding $1 - p_A([1;n], t)$. We observe that $1 - p_A([1;n], t)$ is the probability to spend $< t$ unit of times in the n first states of the infinite chain indexed by $\mathbb{N}$. In particular, the distribution for the total time spent in these n states has for density function a convolution of n exponentials with respective rates $a_{p,p+1}, p = 1 \ldots n$. Furthermore, $a_{p,p+1} \leq \max_{i\in L_p} \sum_{j\neq i} m_{i,j} = \max_{i\in L_p} |m_{i,i}| \leq \max_{i\in L_p} |q_{i,i}| \leq p$ (cf. Condition 1). By stochastic dominance we can assume from now on $a_{p,p+1} = p$ (this can only decrease $p_A([1;n], t)$).

Let X be the sum of n exponential random variables with respective rates $1, 2, \ldots, n$. The density function of X is $\rho_X(t) = n \cdot e^{-t} \cdot (1-e^{-t})^{n-1}$. In particular, we have:

$$\mathbb{P}r[\ X < t\ ] = \int_0^t f_X(s)ds = (1-e^{-t})^n. \tag{6.29}$$

Finally, we note that $1 - p_Q(B_n, t) = p_Q(\mathcal{S} \setminus B_n, t)$, and similarly $1 - p_M(B_n, t) = p_M(\mathcal{S} \setminus B_n, t)$, because the processes X_Q and X_M are non-explosive. By Corollary 3 we so obtain that:

$$\begin{aligned}
1 - p_Q(B_n, t) &= p_Q(\mathcal{S} \setminus B_n, t) && (6.30)\\
&= e^{-t} \cdot p_M(\mathcal{S} \setminus B_n, t) + \int_0^t e^{-s} p_M(\mathcal{S} \setminus B_n, s) ds && (6.31)\\
&= e^{-t} \cdot (1 - p_M(B_n, t)) + \int_0^t e^{-s} (1 - p_M(B_n, s)) ds && (6.32)\\
&\leq e^{-t} \cdot (1 - p_A([1; n], t)) + \int_0^t e^{-s} (1 - p_A([1; n], s)) ds && (6.33)\\
&\leq e^{-t} \mathbb{P}_r[X < t] + \int_0^t e^{-s} \mathbb{P}_r[X < s] ds && (6.34)\\
&= e^{-t}(1 - e^{-t})^n + \frac{1}{n+1}(1 - e^{-t})^{n+1}. && (6.35)
\end{aligned}$$

For $t = \log(n+1)$ we so obtain $1 - p_Q(B_n, t) \leq \frac{2e^{-1}}{n+1}$. ◇

Since, for every $i \notin B_n$ we have by the Property 5 of locality $x_n(i) = 0$, we deduce from Claim 7.1 that we also have $|\mathbb{E}[\bar{x}_n(i)] - p_Q(i, \log(n+1))| \leq \frac{2e^{-1}}{n+1}$. In particular, for C, β large enough we have for every $i \in \mathcal{S}$:

$$|\mathbb{E}[\bar{x}_n(i)] - p_Q(i, \log(n+1))| \leq \frac{K(i)}{n+1} \cdot \sum_{k=1}^{n-1} \frac{1}{k^d}. \tag{6.36}$$

We then need to observe the following:

$$\begin{aligned}
&|\mathbb{E}[\bar{x}_{n+1}(i)] - p_Q(i, \log(n+2))|\\
&\leq \left| \left[(\ \mathbb{E}[\bar{x}_n] - p_Q(\log(n+1))) \left(Id + \frac{1}{n+1} \cdot Q \right) \right]_i \right| && (6.37)\\
&+ \left| \ \mathbb{E}[\bar{x}_{n+1}(i)] - \mathbb{E}[\bar{x}_n(i)] - \frac{1}{n+1} \cdot (\mathbb{E}[x_n] Q)_i \ \right|\\
&+ \left| \ p_Q(i, \log(n+2)) - p_Q(i, \log(n+1)) - \frac{1}{n+1} \cdot (p_Q(\log(n+1)) Q)_i \ \right|.
\end{aligned}$$

We focus on the first error term (the two others will be handled at the end of the proof, with Lemmas 21 and 24, respectively). Specifically, we prove that the following small contraction occurs:

Claim 7.2. *For every $i \in B_{n+1}$:*

$$\left| \left[(\ \mathbb{E}[\bar{x}_n] - p_Q(\log(n+1))) \left(Id + \frac{1}{n+1} \cdot Q \right) \right]_i \right| \leq \frac{K(i)}{n+1} \cdot \sum_{t=1}^{n-1} \frac{1}{t^d} \cdot \left(\ 1 - \frac{\lambda + 1}{n+1} \ \right)$$

with λ the constant in Cond. 3.

Proof. Let $i \in B_{n+1}$ be fixed. Recall that we have:

$$\left[(\mathbb{E}[\bar{x}_n] - p_Q(\log(n+1))) \left(Id + \frac{1}{n+1} \cdot Q \right) \right]_i \tag{6.38}$$

$$= \left(1 - \frac{|q_{i,i}|}{n+1} \right) \cdot (\mathbb{E}[\bar{x}_n(i)] - p_Q(i, \log(n+1))) \tag{6.39}$$

$$+ \frac{1}{n+1} \cdot \sum_{j \neq i} q_{j,i} \cdot (\mathbb{E}[\bar{x}_n(j)] - p_Q(j, \log(n+1))). \tag{6.40}$$

This can be rewritten by only using the coefficients $m_{i,j}$, as follows:

$$(1 - \frac{|q_{i,i}|}{n+1}) \cdot (\mathbb{E}[\bar{x}_n(i)] - p_Q(i, \log(n+1))) \tag{6.41}$$

$$+ \frac{1}{n+1} \cdot \sum_{j \neq i} q_{j,i} \cdot (\mathbb{E}[\bar{x}_n(j)] - p_Q(j, \log(n+1))) \tag{6.42}$$

$$= (1 - \frac{|m_{i,i}| + \sum_{j \neq i} \mu_j}{n+1}) \cdot (\mathbb{E}[\bar{x}_n(i)] - p_Q(i, \log(n+1))) \tag{6.43}$$

$$+ \frac{1}{n+1} \cdot \sum_{j \neq i} (m_{j,i} + \mu_i) \cdot (\mathbb{E}[\bar{x}_n(j)] - p_Q(j, \log(n+1))) \tag{6.44}$$

$$= (1 - \frac{|m_{i,i}| + 1 - \mu_i}{n+1}) \cdot (\mathbb{E}[\bar{x}_n(i)] - p_Q(i, \log(n+1))) \tag{6.45}$$

$$+ \frac{1}{n+1} \cdot \sum_{j \neq i} m_{j,i} \cdot (\mathbb{E}[\bar{x}_n(j)] - p_Q(j, \log(n+1))) \tag{6.46}$$

$$+ \frac{\mu_i}{n+1} \cdot \sum_{j \neq i} (\mathbb{E}[\bar{x}_n(j)] - p_Q(j, \log(n+1))) \tag{6.47}$$

$$= (1 - \frac{|m_{i,i}| + 1 - \mu_i}{n+1}) \cdot (\mathbb{E}[\bar{x}_n(i)] - p_Q(i, \log(n+1))) \tag{6.48}$$

$$+ \frac{1}{n+1} \cdot \sum_{j \neq i} m_{j,i} \cdot (\mathbb{E}[\bar{x}_n(j)] - p_Q(j, \log(n+1))) \tag{6.49}$$

$$+ \frac{\mu_i}{n+1} \cdot (1 - \mathbb{E}[\bar{x}_n(i)] - (1 - p_Q(i, \log(n+1)))) \tag{6.50}$$

$$= \left(1 - \frac{|m_{i,i}| + 1}{n+1} \right) \cdot (\mathbb{E}[\bar{x}_n(i)] - p_Q(i, \log(n+1))) \tag{6.51}$$

$$+ \frac{1}{n+1} \cdot \sum_{j \neq i} m_{j,i} \cdot (\mathbb{E}[\bar{x}_n(j)] - p_Q(j, \log(n+1))) \tag{6.52}$$

$$+ \frac{\mu_i}{n+1} \cdot (\mathbb{E}[\bar{x}_n(i)] - p_Q(i, \log(n+1))) \tag{6.53}$$

$$+ \frac{\mu_i}{n+1} \cdot (p_Q(i, \log(n+1)) - \mathbb{E}[\bar{x}_n(i)]) \tag{6.54}$$

$$= \left(1 - \frac{|m_{i,i}| + 1}{n+1} \right) \cdot (\mathbb{E}[\bar{x}_n(i)] - p_Q(i, \log(n+1))) \tag{6.55}$$

$$+ \frac{1}{n+1} \cdot \sum_{j \neq i} m_{j,i} \cdot (\mathbb{E}[\bar{x}_n(j)] - p_Q(j, \log(n+1))). \tag{6.56}$$

Note that we implicitly used the fact that X_Q is non-explosive and so, $\sum_k p_Q(k,t) = 1$, in order to derive this above equation. Furthermore, since $|m_{i,i}| + 1 \leq n + 1$ for

every $i \in B_{n+1}$ (Condition 1), all coefficients of the above operator are nonegative. We so obtain:

$$\left| \left[(\ \mathbb{E}[\bar{x}_n] - p_Q(\log(n+1))) \left(Id + \frac{1}{n+1} \cdot Q \right) \right]_i \right| \tag{6.57}$$

$$\leq \left(1 - \frac{|m_{i,i}|+1}{n+1} \right) \cdot |\ \mathbb{E}[\bar{x}_n(i)] - p_Q(i, \log(n+1))\ | \tag{6.58}$$

$$+ \frac{1}{n+1} \cdot \sum_{j \neq i} m_{j,i} \cdot |\ \mathbb{E}[\bar{x}_n(j)] - p_Q(j, \log(n+1))\ |\,. \tag{6.59}$$

By the induction hypothesis:

$$\left| \left[(\ \mathbb{E}[\bar{x}_n] - p_Q(\log(n+1))) \left(Id + \frac{1}{n+1} \cdot Q \right) \right]_i \right| \tag{6.60}$$

$$\leq \left(1 - \frac{|m_{i,i}|+1}{n+1} \right) \cdot \frac{K(i)}{n+1} \cdot \sum_{t=1}^{n-1} \frac{1}{t^d} \tag{6.61}$$

$$+ \frac{1}{n+1} \cdot \sum_{j \neq i} m_{j,i} \cdot \frac{K(j)}{n+1} \cdot \sum_{t=1}^{n-1} \frac{1}{t^d}. \tag{6.62}$$

Observe that for every $1 \leq p \leq n+1$, we have if $i \in B_p \setminus B_{p-1}$:

$$\sum_{j \neq i} m_{j,i} K(j) = C(p+1+\beta)^{1+d} \cdot \sum_{j \notin B_p} m_{j,i} \tag{6.63}$$

$$+ C(p-1+\beta)^{1+d} \cdot \sum_{j \in B_{p-1}} m_{j,i} \tag{6.64}$$

$$+ C(p+\beta)^{1+d} \cdot \sum_{j \in B_p \setminus (B_{p-1} \cup \{i\})} m_{j,i} \tag{6.65}$$

$$= K(i) \cdot \left(1 + \frac{1}{p+\beta} \right)^{1+d} \cdot \sum_{j \notin B_p} m_{j,i} \tag{6.66}$$

$$+ K(i) \cdot \left(1 - \frac{1}{p+\beta} \right)^{1+d} \cdot \sum_{j \in B_{p-1}} m_{j,i} \tag{6.67}$$

$$+ K(i) \cdot \sum_{j \in B_p \setminus (B_{p-1} \cup \{i\})} m_{j,i} \tag{6.68}$$

Furthermore, by Taylor-Lagrange inequality we have:

$$\left| \left(1 + \frac{1}{p+\beta} \right)^{1+d} - 1 - \frac{1+d}{p+\beta} \right| \leq \frac{d(1+d)}{2(p+\beta)^2} \tag{6.69}$$

and in the same way:

$$\left| \left(1 - \frac{1}{p+\beta} \right)^{1+d} - 1 + \frac{1+d}{p+\beta} \right| \leq \frac{d(1+d)}{2(p+\beta)^2}. \tag{6.70}$$

We so obtain the following:

$$\sum_{j\neq i} m_{j,i}K(j) \;-\; K(i)\cdot\sum_{j\neq i} m_{j,i} \tag{6.71}$$

$$\leq K(i)\cdot\frac{1+d}{p+\beta}\cdot\left(\sum_{j\notin B_p} m_{j,i} \;-\; \sum_{j\in B_{p-1}} m_{j,i}\right) \tag{6.72}$$

$$+\, K(i)\cdot\frac{d(1+d)}{2(p+\beta)^2}\cdot\sum_{j\neq i} m_{j,i}. \tag{6.73}$$

The first error term above is non-positive, that follows from Condition 2. As we can also deduce from Condition 1 that $|m_{i,i}| \leq p+\mathcal{O}(1)$, the second error term is smaller than $\lambda K(i)$ for some choice of β large enough. Overall, we get from Condition 3 the following chain of inequalities:

$$\sum_{j\neq i} m_{j,i}K(j) \leq K(i)\cdot\left(\sum_{j\neq i} m_{j,i}+\lambda\right) \leq K(i)\cdot(|m_{i,i}|-\lambda). \tag{6.74}$$

We can now conclude the claim, as follows:

$$\left|\left[(\,\mathbb{E}[\bar{x}_n]-p_Q(\log(n+1)))\left(Id+\frac{1}{n+1}\cdot Q\right)\right]_i\right| \tag{6.75}$$

$$\leq\left(1-\frac{1}{n+1}\right)\cdot\frac{K(i)}{n+1}\cdot\sum_{t=1}^{n-1}\frac{1}{t^d} \tag{6.76}$$

$$+\,\frac{1}{n+1}\cdot\sum_{j} m_{j,i}\cdot\frac{K(j)}{n+1}\cdot\sum_{t=1}^{n-1}\frac{1}{t^d} \tag{6.77}$$

$$\leq\frac{K(i)}{n+1}\cdot\sum_{t=1}^{n-1}\frac{1}{t^d}-(\lambda+1)\frac{K(i)}{(n+1)^2}\cdot\sum_{t=1}^{n-1}\frac{1}{t^d}. \tag{6.78}$$

$\diamond$

By combining the above Claim 7.2 with Lemmas 21 and 24, we obtain for n large enough:

$$|\mathbb{E}[\bar{x}_{n+1}(i)]-p_Q(i,\log(n+2))| \tag{6.79}$$

$$\leq\left(\,1-\frac{\lambda+1}{n+1}\,\right)\cdot\frac{K(i)}{n+1}\cdot\sum_{t=1}^{n-1}\frac{1}{t^d} \tag{6.80}$$

$$+\,c'\cdot\left(\frac{|m_{i,i}|}{n}\right)^{1+d}+o\left(\,\frac{1}{(n+1)^2}\,\right) \tag{6.81}$$

$$\leq\left(1+\frac{1}{n+1}\right)\cdot\left(\,1-\frac{\lambda+1}{n+1}\,\right)\cdot\frac{K(i)}{n+2}\cdot\sum_{t=1}^{n-1}\frac{1}{t^d} \tag{6.82}$$

$$+\,c'\cdot\left(\frac{|m_{i,i}|}{n}\right)^{1+d}+o\left(\,\frac{1}{(n+1)^2}\,\right) \tag{6.83}$$

$$\tag{6.84}$$

$$\leq \left(1 - \frac{\lambda}{n+1} - \frac{\lambda+1}{(n+1)^2} \right) \cdot \frac{K(i)}{n+2} \cdot \sum_{t=1}^{n-1} \frac{1}{t^d} \tag{6.85}$$

$$+ c' \cdot \left(\frac{|m_{i,i}|}{n} \right)^{1+d} + o\left(\frac{1}{(n+1)^2} \right) \tag{6.86}$$

$$< \left[\frac{K(i)}{n+2} \cdot \sum_{t=1}^{n-1} \frac{1}{t^d} + c' \cdot \left(\frac{|m_{i,i}|}{n} \right)^{1+d} \right] \tag{6.87}$$

$$- \left[\frac{\lambda K(i)}{(n+2)^2} \cdot \sum_{t=1}^{n-1} \frac{1}{t^d} + o\left(\frac{1}{(n+1)^2} \right) \right] \tag{6.88}$$

$$< \frac{K(i)}{n+2} \cdot \sum_{t=1}^{n-1} \frac{1}{t^d} + c' \cdot \left(\frac{|m_{i,i}|}{n} \right)^{1+d}. \tag{6.89}$$

Thus, for C large enough:

$$\frac{K(i)}{n+2} \cdot \sum_{t=1}^{n-1} \frac{1}{t^d} + c' \cdot \left(\frac{|m_{i,i}|}{n} \right)^{1+d} \leq \frac{K(i)}{n+2} \cdot \sum_{t=1}^{n} \frac{1}{t^d}. \tag{6.90}$$

This achieves proving the induction hypothesis for $n+1$. □

6.4 A truncated process

Theorem 7 shows the link between the normalized mean state distribution and the stationary distribution of some Markov process. The following is a classical technique in order to define such a distribution (if it exists) as the limit of a sequence of stationary distributions for some *finite-state* Markov processes – that are obtained by truncation of the state space, and simpler to compute. Specifically, we recall that for a scale-invariant process $(\mathcal{S}, m, \mu)$ – as defined in Sec. 6.2.1 – we associate some transition-rate matrix Q (cf. Definition 5). Furthermore, for the Markov process defined by Q, there is a natural *rebirth process* with unit rate, to the initial distribution μ. So, we can write $Q = R + M$ with M being the rate matrix of the process without a reset and R corresponding to the reset. By Lemma 20, the unique invariant measure π of Q is defined as $\int_{t\in[0,+\infty]} e^{-t} \cdot p_M(t)dt$. Note that we will reuse this decomposition technique in Sec. 6.5 in order to compute closed formulas for the stationary distributions in some simple cases. We are thus left to study the properties of the state distribution $p_M(t)$ at time t *conditioned on the fact that there has been no reset.*

Intuitively, what we aim at avoiding are the pathological cases where, starting from some initial distribution of finite support, we may reach a state arbitrarily far away in finite time with positive probability. This seemingly weaker form of *non-explosiveness* can happen if, for instance, $\mathcal{S} = \mathbb{N}$ and $\forall i,\ m_{i,i+1} = i^2$. So we shall assume that the process defined by M does not escape toward infinity too fast. This shall allow us first to ensure that the invariant measure π is always a stationary distribution for Q, and then to *truncate* this infinite chain to compute approximations of the stationary distribution.

Definition 6 (non-escaping process)**.** *We say that a process is non-escaping if there exists a function $\rho(\cdot,\cdot)$ increasing in t such that $\forall t, \varepsilon, p_M\left(B_{\rho(t,\varepsilon)}, t\right) > 1-\varepsilon$. Equivalently, the process with no reset remains in a finite ball (whose radius depends on the time elapsed) with high probability*[†]*.*

This non-escaping property is sufficient to ensure the existence of a stationary distribution, as we prove next.

Proposition 1. *If the process is non-escaping, then π is a stationary distribution for Q.*

Proof. By Lemma 20 we already know that π is an invariant measure. Therefore, in order to prove the result, it suffices to prove that $||\pi||_1 = 1$ (*i.e.*, there is no mass escaped toward infinity). For that, we prove that for any $\epsilon > 0$, we have $||\pi||_1 = ||\int_{t\in[0,+\infty]} e^{-t}\cdot p_M(t)dt||_1 > 1-\varepsilon$. Indeed, let t_1 be such that $e^{-t1} \leq \varepsilon/4$. Then, $||\pi - \int_0^{t_1} e^{-t}\cdot p_M(t)dt||_1 \leq \varepsilon/4$. Moreover, according to Definition 6, there is a finite $\rho_1 = \rho(t_1, \varepsilon/4)$ such that, for any $t \leq t_1$, we have $p_M\left(B_{\rho_1}, t\right) > 1-\varepsilon/4$. This implies, as desired, that we have $||\int_0^{t_1} e^{-t}\cdot p_M(t)dt||_1 > (1-\varepsilon/4)\int_0^{t_1} e^{-t}dt > (1-\varepsilon/4)^2 \approx 1-\varepsilon/2$, and so $||\pi||_1 > 1-3\varepsilon/4$. □

In order to approximate the stationary distribution, we shall use a truncated process that evolves in a finite state space and for which finding the stationary distribution is about determining the Perron Frobenius eigenvector of a stochastic matrix. Informally speaking, the truncated process is simply the process restricted to a ball of some radius.

Definition 7. *For $Q = R + M$, its* truncated process at radius ρ*, denoted $R + M_\rho$, is defined on the finite state-space $\mathcal{S}_\rho = \mathcal{S} \cap B_\rho$* [‡] *by the transition-rate matrix $M_\rho = (m^\rho_{i,j})_{i,j\in\mathcal{S}_\rho}$ such that:*

$$\begin{cases} m^\rho_{i,j} = m_{i,j} \text{ if } i \neq j \\ m^\rho_{i,i} = -\sum_{j\neq i} m^\rho_{i,j}. \end{cases}$$

In particular, all transitions $i \to j$ with $i \in \mathcal{S}_\rho, j \notin \mathcal{S}_\rho$ are replaced by loops with the same rate.

Let π_ρ be the stationary distribution of $R + M_\rho$. We prove under the following stronger condition that we can use these finite distributions in order to approximate π at an arbitrary precision.

Definition 8 (strongly non-escaping process)**.** *We say that the process defined by M is strongly non-escaping if for any $t > 0$, there exists a function $\rho'(t,\varepsilon)$ increasing in t such that $\mathbb{P}r[\exists t' < t \text{ s.t. } X_M(t_1+t') \notin B_{\rho'(t,\varepsilon)} \mid X_M(t_1)] < \varepsilon$.*

Proposition 2. *If the process is strongly non-escaping, then for every t, ε we have $||\pi - \pi_{\rho'(t,\varepsilon)}||_1 \leq 2(e^{-t}+\varepsilon)$.*

[†]We may choose $\mathcal{S} = \mathbb{N}^k$ and the balls according to the ℓ_1-norm for our study, however the arguments are quite general and they would apply for other norms.

[‡]We assume that ρ is large enough so that all initial states are in S_ρ, *i.e.*, $\mu(\mathcal{S}_\rho) = 1$.

Proof. By Lemma 20, $||\pi - p_Q(t)||_1 \leq e^{-t}$, and in the same way $||\pi_{\rho'(t,\varepsilon)} - p_{R+M_{\rho'(t,\varepsilon)}}(t)||_1 \leq e^{-t}$. Finally, we observe that $p_{R+M_{\rho'(t,\varepsilon)}}(t) = p_Q(t) \mid \forall t' \leq t, X_Q(t') \in B_{\rho'(t,\varepsilon)}$. Equivalently, the two processes are the same conditioned on the fact that we did not leave the ball $B_{\rho'(t,\varepsilon)}$.In particular, $p_{R+M_{\rho'(t,\varepsilon)}}(t) + \varepsilon \geq p_Q(t) > (1-\varepsilon) \cdot p_{R+M_{\rho'(t,\varepsilon)}}(t)$ since we assume the process to be strongly non-escaping. This implies $||p_Q(t) - p_{R+M_{\rho'(t,\varepsilon)}}(t)||_1 \leq 2\varepsilon$. □

6.5 Computation of the distributions in some practical cases

In this section, we illustrate our method, presented in Sec. 6.3, on some classic one-dimensional and two-dimensional models. In particular, we provide the joint distribution of the in- and out-degrees of the model of [9] which was not known before.

6.5.1 One-dimensional state

We first consider a situation in which the set of states is $\mathbb{N}$, with a unique entering state 1 ($\mu(\{1\}) = 1$). Indeed, we observe that almost all previous studies on PA models, and their relation to Markov theory, focused on this one-dimensional state case. Since the only possible transition from the state i is to the state $i+1$, we now call $m_i \overset{def}{=} m_{i,i+1}$ for better visibility. We denote by $\pi = (S_i)_{i\in\mathbb{N}}$ the stationary distribution (if it exists). It satisfies the following recurrence equation:

$$(m_{i+1}+1)S_{i+1} = m_i S_i \to S_{i+1} = \frac{m_i}{m_{i+1}} \times \frac{1}{1+\frac{1}{m_{i+1}}} \times S_i. \tag{6.91}$$

It follows that S_i can be expressed as:

$$S_i = S_{i_0} \times \frac{m_{i_0}}{m_i} \times \prod_{j=i_0}^{i-1} \frac{1}{1+\frac{1}{m_j}}. \tag{6.92}$$

The above partial product converges if and only if the following series converges:

$$\sum_j \ln\left(\frac{1}{1+\frac{1}{m_j}}\right) = -\sum_j \ln\left(1+\frac{1}{m_j}\right). \tag{6.93}$$

In the case $m_j \gg 1$, $\ln\left(1+\frac{1}{m_j}\right) \sim \frac{1}{m_j}$, and we are left to study the convergence of $\sum_j \frac{1}{m_j}$.

First case, $\sum_j \frac{1}{m_j}$ converges. In this case, the process is escaping and the system of stationary equations is inconsistent. Indeed, if $\sum_j \frac{1}{m_j}$ converges, then $S_i = \Theta(\frac{1}{m_i})$. The latter implies that $\sum_i S_i$ converges, that proves that the stationary equations do have a solution. But something unusual happens: the probability $\frac{m_i}{1+m_i} \sim 1 - \frac{1}{m_i}$ can be interpreted as the probability that the transition from i to $i+1$ happens before a reset transition. Therefore,

$$\prod_{j\geq i}(1-\frac{1}{m_j}) \sim \exp(-\sum_{j\geq i}\frac{1}{m_j}) \tag{6.94}$$

determines the nature of the probability to reach $+\infty$ when at i. When $\sum_j \frac{1}{m_j}$ converges, this probability becomes greater than 0 and there is no standard stationary distribution. Then, the system of stationary equations is also inconsistent. Indeed, we must have

$$S_{i_0}(m_{i_0}+1) = 1 \to S_{i_0} = \frac{1}{1+m_{i_0}}, \tag{6.95}$$

and so the series

$$\sum_{i \geq i_0} \frac{1}{m_i+1} \times \prod_{j=i_0}^{i-1} \frac{1}{1+\frac{1}{m_j}} \tag{6.96}$$

should sum to 1, which is not the case.

Example: Set $i_0 = 1$ and $m_i = i^2$. We have

$$S_i = \frac{S_{i_0}}{i^2+1} \prod_{j=1}^{i} \frac{1}{1+\frac{1}{i^2}}.$$

Since

$$\frac{\sin z}{z} = \prod_j \left(1 - \frac{z^2}{\pi^2 j^2}\right),$$

we finally obtain that [§]:

$$S_i \sim \frac{2S_{i_0}}{i^2+1} \frac{\sqrt{-1}\pi}{\sin \sqrt{-1}\pi} \sim \frac{2\pi S_{i_0}}{\mathrm{e}^{\pi} - \mathrm{e}^{-\pi}} \cdot \frac{1}{i^2}.$$

But this does not correspond to an actual stationary distribution, and the series S_i does not sum to 1 when we set $S_{i_0=1} = \frac{1}{2}$.

Second case, $\sum_j \frac{1}{m_i}$ diverges. If $\sum_j \frac{1}{m_i}$ diverges, when m is an integrable increasing function with $\lim_{x\to\infty} m_x = \infty$, we let $M_i = \int_{i_0}^{i} \mathrm{d}t/m_t$. Then,

$$\left(\sum_{j=i_0}^{i} \frac{1}{m_j}\right) - \frac{1}{m_{i_0}} \leq M_i \leq \left(\sum_{j=i_0}^{i} \frac{1}{m_j}\right) - \frac{1}{m_i}.$$

Therefore, we have

$$S_i = \Theta\left(\frac{1}{m_i} \exp\left(-\sum_{j\in[i_0,i]} \frac{1}{m_j}\right)\right) = \Theta\left(\frac{\exp\left(-M_i\right)}{m_i}\right),$$

and the stationary distribution exists if and only if $\sum_i \exp\left(-M_i\right)/m_i$ converges.

Example: If we let $i_0 = 1$ and $m_i = ai + b$, we find that

$$\frac{\exp\left(-M_i\right)}{m_i} = \exp\left(-\frac{1}{a}\ln(ai+b)\right)/(ai+b) = \Theta\left(i^{-\left(1+\frac{1}{a}\right)}\right).$$

Therefore $S_i = \Theta(i^{-(1+1/a)})$, which is the usual power law.

[§]We use here the notation $\mathrm{i} = \sqrt{-1}$ in order to keep clear the distinction between our parameter and the imaginary number.

Note also that, if we denote $C_i = \sum_{j>i} S_i = 1 - \sum_{j\leq i} S_i$, the stability equations imply that:

$$C_i = 1 - \sum_{j\leq i} S_i = m_i S_i = m_i(\sum_{j\leq i} S_j - \sum_{j\leq i-1} S_j) = m_i(C_{i-1} - C_i).$$

Thus,

$$C_i = \frac{m_i}{m_i+1} C_{i-1} \to C_i = C_{i_0} \prod_{j\in[i_0,i]} \frac{m_j}{m_j+1}.$$

And so, C_i behaves like $\exp(-\sum_j \frac{1}{m_j})$ and it tends towards 0 only if $\sum_j \frac{1}{m_j} \to +\infty$.

In this simple case, one can derive an exact and closed formula for the S_i's of a generalization of the Barabási-Albert model using the Gamma function. Namely:

Proposition 3 (Folk)**.** *For $m_i = ai + b, a \in]0, +\infty[$ then, $\forall i \geq i_0$:*

$$S_i = S_{i_0} \cdot \frac{\Gamma\left(i + \frac{b}{a}\right)\Gamma\left(i_0 + 1 + \frac{b+1}{a}\right)}{\Gamma\left(i_0 + \frac{b}{a}\right)\Gamma\left(i + 1 + \frac{b+1}{a}\right)} \sim_\infty \left[S_{i_0} \cdot \frac{\Gamma\left(i_0 + 1 + \frac{b+1}{a}\right)}{\Gamma\left(i_0 + \frac{b}{a}\right)}\right] \cdot i^{-\left(1+\frac{1}{a}\right)}.$$

Moreover, the process with rate m_i and return rate 1 to i_0 is stable and admits S_i as stationary distribution.

Note also that $S(i) \sim_\infty c(a,b) i^{-(1+\frac{1}{a})}$, where $c(a,b)$ is a suitable constant.

Proof. We start from Equation 6.91 when one sets $m_i = ai + b$:

$$\forall i \geq i_0,\ S_{i+1} = \frac{ai+b}{a(i+1)+b+1} S_i.$$

By induction we get:

$$\begin{aligned}\forall i \geq i_0,\ S_i &= S_{i_0} \cdot \Pi_{j=i_0}^{i-1} \frac{aj+b}{a(j+1)+b+1} \\ &= S_{i_0} \cdot \Pi_{j=0}^{i-1-i_0} \frac{aj + (b+ai_0)}{aj + (b+1+a(i_0+1))}.\end{aligned}$$

Dividing by a the numerators and denominators, we get:

$$\forall i \geq i_0,\ S_i = S_{i_0} \cdot \frac{\Pi_{j=0}^{i-1-i_0}(j + \frac{b+ai_0}{a})}{\Pi_{j=0}^{i-1-i_0}(j + \frac{b+1+a(i_0+1)}{a})}.$$

We can use here the Γ function since

$$\Pi_{k=0}^{n}(k+\beta) = \frac{\Gamma(n+1+\beta)}{\Gamma(\beta)}.$$

This leads to:

$$S_i = S_{i_0} \cdot \frac{\Gamma\left(i - i_0 + \frac{b+ai_0}{a}\right)/\Gamma\left(\frac{b+ai_0}{a}\right)}{\Gamma\left(i - i_0 + \frac{b+1+a(i_0+1)}{a}\right)/\Gamma\left(\frac{b+1+a(i_0+1)}{a}\right)}.$$

This simplifies further as:

$$S_i = S_{i_0} \cdot \frac{\Gamma\left(i + \frac{b}{a}\right)\Gamma\left(i_0 + 1 + \frac{b+1}{a}\right)}{\Gamma\left(i_0 + \frac{b}{a}\right)\Gamma\left(i + 1 + \frac{b+1}{a}\right)}.$$

To estimate S_i, we use Stirling's formula that states that $\Gamma(x) \sim_\infty \left(\frac{x}{e}\right)^x \sqrt{\frac{2\pi}{x}}$. We get:

$$\Gamma\left(i+\frac{b}{a}\right) \sim_\infty \sqrt{2\pi}\left(i+\frac{b}{a}\right)^{i+\frac{b}{a}-\frac{1}{2}} e^{-i-\frac{b}{a}}$$
$$\sim_\infty \sqrt{2\pi}e^{-\frac{b}{a}} \cdot \left(i+\frac{b}{a}\right)^{i} \cdot i^{\frac{b}{a}-\frac{1}{2}} \cdot e^{-i},$$

$$\Gamma\left(i+1+\frac{b+1}{a}\right) \sim_\infty \sqrt{2\pi}\left(i+1+\frac{b+1}{a}\right)^{i+1+\frac{b+1}{a}-\frac{1}{2}} e^{-i-1-\frac{b+1}{a}}$$
$$\sim_\infty \sqrt{2\pi}e^{-1-\frac{b+1}{a}} \cdot \left(i+1+\frac{b+1}{a}\right)^{i} \cdot i^{\frac{1}{2}+\frac{b+1}{a}} \cdot e^{-i}.$$

Combining both, we get:

$$\frac{\Gamma\left(i+\frac{b}{a}\right)}{\Gamma\left(i+1+\frac{b+1}{a}\right)} \sim_\infty e^{1+\frac{1}{a}} \cdot \left(1-\frac{1+\frac{1}{a}}{i+1+\frac{b+1}{a}}\right)^{i} \cdot i^{-\left(1+\frac{1}{a}\right)} \sim_\infty i^{-\left(1+\frac{1}{a}\right)}.$$

Altogether, we get the following asymptotic for S_i:

$$S_i \sim_\infty \left[S_{i_0} \cdot \frac{\Gamma\left(i_0+1+\frac{b+1}{a}\right)}{\Gamma\left(i_0+\frac{b}{a}\right)}\right] \cdot i^{-\left(1+\frac{1}{a}\right)}.$$

□

6.5.2 k-dimensional state

In this Section, we prove some results on the k-dimensional state case, in particular the joint in-out degree distribution of the Bollobás et al. model [9].

Formally, we denote $e_l, l \in [k]$ the l-th vector of the canonical base and we suppose that the transitions are all of the form[¶] $s \to s + \sigma e_l,\ \sigma \in \mathbb{Z}, l \in [k]$ and that the rate of the transition $\Gamma_{l,\sigma} s \to s + \sigma e_l$ depends only on σ and $s_l = s.e_l$. The process can then be considered as k parallel independent processes each taking place in one dimension. The only correlation between these processes is the *common reset*. Indeed, when this event happens, all the dimensions reset at once to some initial state.

Proposition 4. *For a process in dimension k for which the transition rates are independent, we have:*

$$S_s = \sum_{s_0 \in \mathbb{N}^k} \mu(s_0) \int_{T \geq 0} \prod_{l \in [k]} \mathbb{P}r_l[s_l \mid T, s_0.e_l] e^{-T} dT,$$

where $\mathbb{P}r_l[x \mid T, y]$ denotes the probability for the process in dimension l to be in state x after T steps and starting from state y.

[¶] σ is here generic, so we can have many different "shifts" per dimension with varying rates.

Proof. Let us define the *survival time* T as the time elapsed without any reset to an initial state. When (i) the survival time T and (ii) the initial state s_0 on which that last reset happened are fixed, the processes controlling the various directions of the walk become independent. So conditioning on T and s_0, we can express $\mathbb{P}r[s \mid T, s_0]$ as a product:

$$\forall s \in \mathbb{N}^k, \mathbb{P}r[s \mid T, s_0] = \prod_{l\in[k]} \mathbb{P}r_l[s_l.e_l \mid T, s_0.e_l].$$

Integrating over the survival time and the initial state, we have:

$$S_s = \sum_{s_0\in\mathbb{N}^k} \mu(s_0) \int_{T\geq 0} \mathbb{P}r[s \mid T, s_0]e^{-T}dT$$

$$S_s = \sum_{s_0\in\mathbb{N}^k} \mu(s_0) \int_{T\geq 0} \prod_{l\in[k]} \mathbb{P}r[s_l \mid T, s_0.e_l]e^{-T}dT.$$

□

According to Proposition 4, one only has to determine the values of $\mathbb{P}r_l[i \mid T, i_0]$, and, then, to integrate over the survival time. So, determining the stationary distribution reduces to the study of several one-dimensional systems. We now study this later case, a one-dimensional process with a unique reset point i_0 (i.e. $\mu_0(\{i_0\}) = 1$) and with transitions $j \xrightarrow{m(j)} j+1$.

Definition 9. $\{\mathbb{P}r[i \mid t, i_0], j \xrightarrow{m(j)} j+1\}$ *is defined as the probability to be in state* i *after* t *steps, starting from state* i_0*, and with transitions* $j \longrightarrow j+1$ *at rate* $m(j)$*.*

Lemma 25. *For a general dimension* 1 *process with reset at* i_0*, one have* $\forall i > i_0$*:*

$$\{\mathbb{P}r[i \mid t, i_0], j \xrightarrow{m(j)} j+1\} = \int_{t_{i_0},\dots,t_{i-1} \mid \sum_{j\in[i_0,i]} t_j \leq t} \left(\prod_{j\in[i_0,i-1]} m(j)\mathrm{e}^{-m(j)t_j} \right) \mathrm{e}^{-m(i)(t-\sum_{j\in[i_0,i-1]} t_j)} dt_{i_0} \dots dt_{i-1}.$$

Proof. In such a system the time T_j spent in a state $j < i$ follows an exponential law with rate $m(j)$ and so density $m(j)\mathrm{e}^{-m(j)t}$. Moreover, these sojourn times T_j are independent. The distribution law for the time T_i is slightly different since we must use the probability to stay in the state longer than T_i. □

The integral appearing above is, up to a factor, a well-studied case of convolution for which the author in [3] gave a "closed formula". The result is stated in the following lemma:

Lemma 26 ([3])**.** *The convolution of measures with exponential densities* $\mathrm{e}^{-\beta_1}, \dots, \mathrm{e}^{-\beta_n}$ *is:*

$$A_n(t) = \sum_{j\in[0,i]} \frac{e^{-\beta_j t}}{\prod_{l\in[0,i],l\neq j}(\beta_l - \beta_j)}. \tag{6.97}$$

Using the above result enables to go one step further in the computation of $\{\mathbb{P}r[i \mid t, 0], (j \xrightarrow{m(j)} j+1)\}$. We omit the resulting formula here since we cannot perform the integration over the sojourn time in the general case. Nevertheless, in the case of affine rates, we complete the computation in the next section.

The case of one-dimensional affine rates

We now study the case of a process in $\mathbb{N}$ that starts from state i_0 with affine transition rates $j \xrightarrow{aj+b} j+1$ which resets to i_0. We first consider the canonical case of the process $\{0, j \xrightarrow{j+\beta} j+1\}$ instead of $\{0, j \xrightarrow{aj+b} j+1\}$. We will then show that other cases can be reduced to it.

Corollary 4. $\{\mathbb{P}r[i \mid t, 0], j \xrightarrow{j+\beta} j+1\} = \frac{\Gamma(i+\beta)}{i!\Gamma(\beta)} e^{-\beta t}(1-e^{-t})^i$.

Proof. We get from Lemma 25, with $m(j) = j + \beta$:

$$\begin{aligned}
&\{\mathbb{P}r[i \mid t, 0], j \xrightarrow{m(j)} j+1\} \; = \\
&= \; \Pi_{j\in[0,i-1]}(j+\beta) \times \int_{t_0,\ldots,t_{i-1}|\sum_{j\in[0,i-1]} t_j \le t} \mathrm{e}^{-\sum\limits_{j\in[0,i-1]} (j+\beta)t_j} \mathrm{e}^{-(i+\beta)(t-\sum\limits_{j\in[0,i-1]} t_j)} dt_0 \ldots dt_{i-1} \\
&= \; \Pi_{j\in[0,i-1]}(j+\beta) \times \mathrm{e}^{-\beta t} \int_{t_0,\ldots,t_{i-1}|\sum_{j\in[0,i-1]} t_j \le t} \mathrm{e}^{-\sum\limits_{j\in[0,i-1]} jt_j} \mathrm{e}^{-i(t-\sum\limits_{j\in[0,i-1]} t_j)} dt_0 \ldots dt_{i-1} \\
&= \; \tfrac{\Gamma(i+\beta)}{\Gamma(\beta)} e^{-\beta t} I.
\end{aligned}$$

Here $I = \int_{t_0,\ldots,t_{i-1}|\sum_{j\in[0,i-1]t_j} \le t} \mathrm{e}^{-\sum\limits_{j\in[0,i-1]} jt_j} \mathrm{e}^{-i(t-\sum\limits_{j\in[0,i-1]} t_j)} dt_0 \ldots dt_{i-1}$ is the convolution of measures with exponential densities with rates $j, j \in [0, i]$. Using Lemma 26 in the case $\beta_j = j$, we obtain:

$$\begin{aligned}
I &= \sum_{j\in[0,i]} \frac{e^{-jt}}{\prod_{l\in[0,i],l\neq j}(l-j)} \\
&= \sum_{j\in[0,i]} \frac{(-1)^j}{j!(i-j)!} e^{-jt} \\
&= \frac{(1-e^{-t})^i}{i!}. \qquad (6.98)
\end{aligned}$$

It follows that $\{\mathbb{P}r[i \mid t, 0], j \xrightarrow{m(j)} j+1\} = \frac{\Gamma(i+\beta)}{i!\Gamma(\beta)} e^{-\beta t}(1-e^{-t})^i$. □

Now, for general values of a, b we have:

Corollary 5.

$$\{\mathbb{P}r[i \mid t, i_0], j \xrightarrow{aj+b} j+1\} = \frac{\Gamma(i+\frac{b}{a})}{\Gamma(i_0+\frac{b}{a})(i-i_0)!} e^{-(b+ai_0)t}(1-e^{-at})^{i-i_0}.$$

Proof. By shifting the origin, the system $\{i_0, j \xrightarrow{aj+b} j+1\}$ is equivalent to the system $\{0, j \xrightarrow{ai+b'} j+1\}$ with $b' = ai_0 + b$. Then, by rescaling the time by a factor a, we can replace $\{\mathbb{P}r[i \mid T, 0], aj + b'\}$ by $\{\mathbb{P}r[i \mid aT, 0], j + \frac{b'}{a}\}$. It follows that:

$$\begin{aligned}
\{\mathbb{P}r[i \mid t, i_0], j \xrightarrow{aj+b} j+1\} &= \{\mathbb{P}r[i-i_0 \mid t, 0], j \xrightarrow{aj+(b+ai_0)} j+1\} \\
&= \{\mathbb{P}r[i-i_0 \mid at, 0], j \xrightarrow{j+(\frac{b}{a}+i_0)} j+1\}.
\end{aligned}$$

The proof then follows from Corollary 4. □

k-dimensional independent and affine processes.

For the 2-dimensional model of Bollabás et al. [9], there are two initial states $e_0 = (0,1)$ and $e_1 = (1,0)$ to which we return at rates $\gamma = \mu_0(e_0)$ and $\alpha = \mu_0(e_1)$, respectively. We study a more general[‖] k-dimensional system that resets only on the vectors $\{e_l\}_{l\in[k]}$ of the canonical base with reset probability distribution $\mu_0(\{e_l\}) \geq 0, l \in [k], \sum_{l\in[k]} \mu_0(\{e_l\}) = 1$ and that has rate $a_l i_l + b_l$ for the transition $s \overset{a_l s_l + b_l}{\longrightarrow} s + e_l$ in dimension $l \in [k]$.

Proposition 5. *In a k-dimensional system with rates $j \overset{a_l j + b_l}{\longrightarrow} j+1$ and reset measure μ_0, such that $\sum_{l\in[k]} \mu_0(\{e_l\}) = 1$, one has:*

$$S_s = \sum_{m\in[k]} \mu_0(\{e_m\}) A_{m,s} \int_{t\geq 0} H_m(s,t) e^{-t} dt, \tag{6.99}$$

where $A_{m,s}$ is the following constant

$$A_{m,s} = \frac{\Gamma(s_m + \frac{b_m}{a_m})}{\Gamma(\frac{b_m}{a_m}+1)(s_m-1)!} \Pi_{l\neq m} \frac{\Gamma(s_l + \frac{b_l}{a_l})}{\Gamma(\frac{b_l}{a_l}) s_l!}$$

and where $H_m(s,t)$ is defined by

$$H_m(s,t) = e^{-(b_m+a_m)t}(1-e^{-a_m t})^{s_m-1} \Pi_{l\neq m} e^{-b_l t}(1-e^{-a_l t})^{s_l}.$$

Proof. We consider the reset at e_m for some $m \in [k]$. In order to use Corollary 5 for the process in dimension $l \in [k]$, we have to distinguish two slightly different cases, since i_0 is either 0 or 1:

- For a dimension $l \neq m$, the initial state is 0, and so:

$$\{\mathbb{P}r[i = s_l \mid t, 0], j \overset{a_l j + b_l}{\longrightarrow} j+1\} = \frac{\Gamma(i + \frac{b_l}{a_l})}{\Gamma(\frac{b_l}{a_l}) i!} e^{-b_l t} \ (1-e^{-a_l t})^i$$

- For the dimension $l = m$, the initial state is 1, so:

$$\{\mathbb{P}r[i = s_l \mid t, 1], j \overset{a_l j + b_l}{\longrightarrow} j+1\} = \frac{\Gamma(i + \frac{b_l}{a_l})}{\Gamma(\frac{b_m}{a_m}+1)(i-1)!} e^{-(b_m+a_m)t}(1-e^{-a_m t})^{i-1}.$$

Now, using Proposition 4, we get:

$$S_s = \sum_{m\in[k]} \mu_0(\{e_m\}) \int_{T>0} \mathbb{P}r[s|t, e_m] e^{-t} dt,$$

where $\mathbb{P}r[s|t,e_m] = A_{m,s} \times H_m(s,t)$ with

$$\begin{aligned} A_{m,s} &= \frac{\Gamma(s_m + \frac{b_m}{a_m})}{\Gamma(\frac{b_m}{a_m}+1)(s_m-1)!} \Pi_{l\neq m} \frac{\Gamma(s_l + \frac{b_l}{a_l})}{\Gamma(\frac{b_l}{a_l}) s_l!} \\ &\text{and} \\ H_{m,s}(t) &= e^{-(b_m+a_m)t}(1-e^{-a_m t})^{s_m-1} \Pi_{l\in[k], l\neq m} e^{-b_l t}(1-e^{-a_l t})^{s_l}. \end{aligned}$$

□

[‖] Our calculation directly extends to any initial state distribution μ_0.

The function we have to integrate in Proposition 5 is of the form:

$$G(j_0, \dots, j_{k-1}, t) = e^{-ct}\Pi_{h\in[k]}(1 - e^{-d_h t})^{j_h}.$$

We did not manage to get its exact computation**, but we are able to provide the first-order of this integral:

Lemma 27. *Let* $G(j_0, \dots j_{k-1}, t) = e^{-ct}\Pi_{h\in[k]}(1-e^{-d_h t})^{j_h}$ *where* c *and* a_h, b_h, *for* $h \in [k]$ *are positive real numbers. Then,*

$$I = \int_{t>0} G(j_0, j_1, \dots j_{k-1}, t)dt = \frac{\Theta(1)}{\max_{h\in[k]} j_h^{c/d_h}}.$$

Proof. Letting $x = e^{-ct}$ and performing a change of variable, we get

$$I = \frac{1}{c}\int_{x\in[0,1]} (1 - x^{d_0/c})^{j_0}(1 - x^{d_1/c})^{j_1} \cdots (1 - x^{d_m/c})^{j_m} dx.$$

Next, we consider the function $f(x) = (1 - x^a)^j$ and remark that:

(i) $\forall x \in [0, 1]$, $f(x) \leq 1$;

(ii) $\forall x \in [0, \frac{1}{j^{1/a}}]$, $f(x) \geq e^{-1}$;

(iii) Beyond $\frac{1}{j^{1/a}}$, f decreases so sharply that $\int_{x\in[0,1]} f(x)dx = \frac{\Gamma(1+\frac{1}{a})\Gamma(1+j)}{\Gamma(1+\frac{1}{a}+j)} = \Theta(\frac{1}{j^{1/a}})$.

Now, letting $v_{min} = \min_{h\in[m]} \frac{1}{j_h^{c/d_h}}$, we get:

(1) $I \leq v_{min}$ (using (i) and (iii)) and

(2) $I \geq e^{-|[m]|}v_{min}$ (using (ii)).

□

Remark 12. *For seek of simplicity we just provide the first-order of* G. *However, one can probably determine an equivalent using a more accurate saddle method. We stress that first order formulas are already quite tedious to obtain.*

Corollary 6.

$$\int_{t\geq 0} H_m(s, t)e^{-t}dt = \frac{\Theta(1)}{\max_{l\in[k]} s_l^{c/a_l}}, \quad c = 1 + a_m + \sum_{l\in[k]} b_l.$$

Proof. We apply Lemma 27 with $c = 1+a_m+\sum_{l\in[k]} b_l$, $d_h = a_h$ and $j_h = s_h - \delta_{h,m} \sim_{s_h\gg 1} s_h$, where $\delta_{h,m}$ is the Kronecker symbol. □

With the above estimate we can determine S_s using Proposition 5, just by summing over the dimensions. We next exemplify this process using the Bollobàs Model.

**When $h = 1$ one can easily compute the integral using an integration by parts, that leads to an induction similar to the one of the Γ function. The general case is possibly related to a k-dimensional extension of the Γ function, but we fail to find any reference to this kind of generalisation.

Application to the model of Bollobás et al. [9]

We already mentioned the model of Bollobás et al. in Section 6.2.1. It corresponds to a two-dimensional affine and independent process. We can thus apply Proposition 5.

Proposition 6. *In the model of Bollobás et al. [9], the joint degree distribution is:*

$$S_{i,j} = \begin{cases} \Theta(i^{-\frac{1+a_0+a_1+b_1}{a_0}} j^{\frac{b_1}{a_1}}) & \text{if } i^{-\frac{1}{a_0}} < j^{-\frac{1}{a_1}} \\ \Theta(i^{\frac{b_0}{a_0}} j^{-\frac{1+a_0+a_1+b_0}{a_1}}) & \text{otherwise.} \end{cases}$$

Proof. We saw that in the considered model the state of a node v is the vector $(d^-(v), d^+(v))$, with two initial states $e_0 = (1,0)$ and $e_1 = (0,1)$ to which we return at rates $\mu_0(e_0) = \frac{\alpha}{\alpha+\gamma}$ and $\mu_0(e_1) = \frac{\gamma}{\alpha+\gamma}$, respectively. The transitions are: $s \longrightarrow s+e_0 : m_{(i,j),(i+1,j)} = \frac{(\alpha+\beta)}{(1-\beta)} \cdot \frac{i+\delta_{in}}{1+(1-\beta)\cdot\delta_{in}}$, and $s \longrightarrow s+e_1 : m_{(i,j),(i,j+1)} = \frac{(\gamma+\beta)}{(1-\beta)} \cdot \frac{j+\delta_{out}}{1+(1-\beta)\cdot\delta_{out}}$. In our context, it corresponds to the parameters:

a_0	$\frac{(\alpha+\beta)}{(1-\beta)(1+(1-\beta)\cdot\delta_{in})}$	b_0	$\frac{(\alpha+\beta)\delta_{in}}{(1-\beta)(1+(1-\beta)\cdot\delta_{in})}$
a_1	$\frac{(\gamma+\beta)}{(1-\beta)(1+(1-\beta)\cdot\delta_{out})}$	b_1	$\frac{(\gamma+\beta)\delta_{out}}{(1-\beta)(1+(1-\beta)\cdot\delta_{out})}$

We can thus use Proposition 5 in this particular case:

$$S_{i,j} = \frac{\alpha}{\alpha+\gamma} A_{0,(i,j)} \int_{t\geq 0} H_0((i,j),t)e^{-t}dt + \frac{\gamma}{\alpha+\gamma} A_{1,(i,j)} \int_{t\geq 0} H_1((i,j),t)e^{-t}dt$$

We have:

$$A_{0,(i,j)} = \frac{\Gamma(i+\frac{b_0}{a_0})}{\Gamma(\frac{b_0}{a_0}+1)(i-1)!} \frac{\Gamma(j+\frac{b_1}{a_1})}{\Gamma(\frac{b_1}{a_1})j!} = \Theta(1) i^{\frac{b_0}{a_0}} j^{\frac{b_1}{a_1}-1}$$
$$A_{1,(i,j)} = \frac{\Gamma(j+\frac{b_1}{a_1})}{\Gamma(\frac{b_1}{a_1}+1)(j-1)!} \frac{\Gamma(i+\frac{b_0}{a_0})}{\Gamma(\frac{b_0}{a_0})i!} = \Theta(1) j^{\frac{b_1}{a_1}} i^{\frac{b_0}{a_0}-1}$$

and

$$H_0((i,j),t) = \Theta(1) \frac{1}{\max(i^{\frac{1+a_0+b_0+b_1}{a_0}}, j^{\frac{1+a_0+b_0+b_1}{a_1}})}$$
$$H_1((i,j),t) = \Theta(1) \frac{1}{\max(i^{\frac{1+a_1+b_0+b_1}{a_0}}, j^{\frac{1+a_1+b_0+b_1}{a_1}})}.$$

Since

$$\max(i^{\frac{1+a_1+b_0+b_1}{a_0}}, j^{\frac{1+a_1+b_0+b_1}{a_1}}) = \begin{cases} i^{\frac{1+a_1+b_0+b_1}{a_0}} & \text{if } i^{-\frac{1}{a_0}} < j^{-\frac{1}{a_1}} \\ j^{\frac{1+a_1+b_0+b_1}{a_1}} & \text{otherwise,} \end{cases}$$

we get

$$S_{i,j} = \begin{cases} \Theta(i^{\frac{b_0}{a_0}} j^{\frac{b_1}{a_1}} (\frac{\alpha}{\alpha+\gamma} i^{-\frac{1+a_0+b_0+b_1}{a_0}} j^{-1} + \frac{\gamma}{\alpha+\gamma} i^{-1-\frac{1+a_1+b_0+b_1}{a_0}})) & \text{if } i^{-\frac{1}{a_0}} < j^{-\frac{1}{a_1}} \\ \Theta(i^{\frac{b_0}{a_0}} j^{\frac{b_1}{a_1}} (\frac{\alpha}{\alpha+\gamma} j^{-1-\frac{1+a_0+b_0+b_1}{a_1}} + \frac{\gamma}{\alpha+\gamma} i^{-1} j^{-\frac{1+a_1+b_0+b_1}{a_1}})) & \text{otherwise.} \end{cases}$$

Using the condition $i^{-\frac{1}{a_0}} < j^{-\frac{1}{a_1}}$, we can neglect one of the terms in each case, which gives:

$$S_{i,j} = \begin{cases} \Theta(i^{-\frac{1+a_0+a_1+b_1}{a_0}} j^{\frac{b_1}{a_1}}) & \text{if } i^{-\frac{1}{a_0}} < j^{-\frac{1}{a_1}} \\ \Theta(i^{\frac{b_0}{a_0}} j^{-\frac{1+a_0+a_1+b_0}{a_1}}) & \text{otherwise.} \end{cases}$$

□

Remark 13. *One can also derive a somewhat "closed" and exact expression for $S_{i,j}$. To that aim, recall that when using Proposition 5 in the two dimensional we had :*

$$S_{i,j} = \frac{\alpha}{\alpha+\gamma} A_{0,(i,j)} \int_{t\geq 0} H_0((i,j),t)e^{-t}dt + \frac{\gamma}{\alpha+\gamma} A_{1,(i,j)} \int_{t\geq 0} H_1((i,j),t)e^{-t}dt$$

with

$$A_{0,(i,j)} = \frac{\Gamma(i+\frac{b_0}{a_0})}{\Gamma(\frac{b_0}{a_0}+1)(i-1)!} \frac{\Gamma(j+\frac{b_1}{a_1})}{\Gamma(\frac{b_1}{a_1})j!} \quad \Big| \quad A_{1,(i,j)} = \frac{\Gamma(j+\frac{b_1}{a_1})}{\Gamma(\frac{b_1}{a_1}+1)(j-1)!} \frac{\Gamma(i+\frac{b_0}{a_0})}{\Gamma(\frac{b_0}{a_0})i!}$$

since $H_{m,s}(t) = e^{-(b_m+a_m)t}(1-e^{-a_m t})^{s_m-1}\Pi_{l\in[k],l\neq m} e^{-b_l t}(1-e^{-a_l t})^{s_l}$ we get :

$$H_0((i,j),t) = e^{-(b_0+b_1+a_0)t}(1-e^{-a_0 t})^{i-1}(1-e^{-a_1 t})^{j}$$
$$H_1((i,j),t) = e^{-(b_0+b_1+a_1)t}(1-e^{-a_1 t})^{j-1}(1-e^{-a_0 t})^{i}$$

We can then expand the formula using the binomial expansion to get :

$$H_0((i,j),t) = e^{-(b_0+b_1+a_0)t} \sum_{l\in[i-1]} (-1)^l \binom{i-1}{l} e^{-a_0 l t} \times \sum_{k\in[j]} (-1)^k \binom{j}{k} e^{-a_1 k t}$$

$$H_0((i,j),t) = \sum_{l\in[i-1],k\in[j]} (-1)^{l+k} \binom{i-1}{l}\binom{j}{k} e^{-(a_0 l + a_1 k + b_0 + b_1 + a_0)t}$$

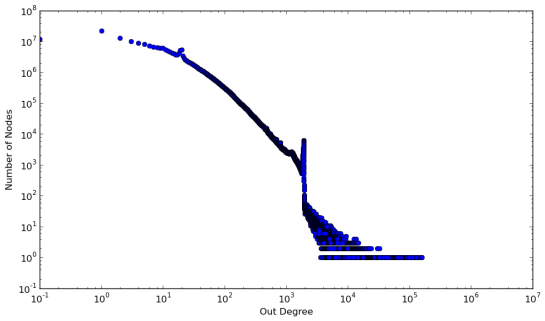
$$\int_{t>0} H_0((i,j),t)e^{-t} = \sum_{l\in[i-1],k\in[j]} (-1)^{l+k} \binom{i-1}{l}\binom{j}{k} \frac{1}{1+a_0 l + a_1 k + b_0 + b_1 + a_0}$$

Symmetrically we have

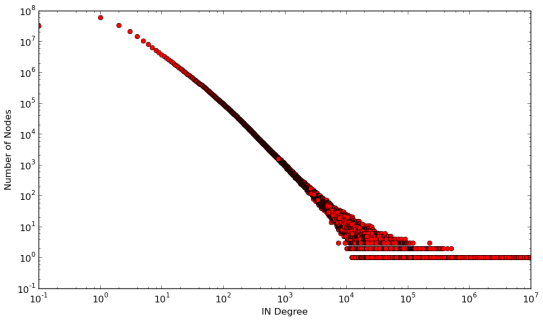
$$\int_{t>0} H_1((i,j),t)e^{-t} = \sum_{l\in[j-1],k\in[i]} (-1)^{l+k} \binom{j-1}{l}\binom{i}{k} \frac{1}{1+a_1 l + a_0 k + b_1 + b_0 + a_1}. \tag{6.100}$$

The global formula is then :

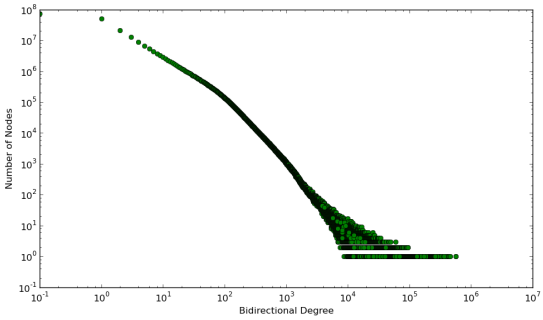
$$S_{i,j} =$$
$$\frac{\alpha}{\alpha+\gamma} \frac{\Gamma(i+\frac{b_0}{a_0})}{\Gamma(\frac{b_0}{a_0}+1)} \frac{\Gamma(j+\frac{b_1}{a_1})}{\Gamma(\frac{b_1}{a_1})} \sum_{l\in[i-1],k\in[j]} \frac{(-1)^{l+k}}{(i-1-l)!(j-k)!l!k!} \frac{1}{1+a_0 l + a_1 k + b_0 + b_1 + a_0}$$
$$+ \frac{\gamma}{\alpha+\gamma} \frac{\Gamma(j+\frac{b_1}{a_1})}{\Gamma(\frac{b_1}{a_1}+1)} \frac{\Gamma(i+\frac{b_0}{a_0})}{\Gamma(\frac{b_0}{a_0})} \sum_{l\in[j-1],k\in[i]} \frac{(-1)^{l+k}}{(j-1-l)!(i-k)!l!k!} \frac{1}{1+a_1 l + a_0 k + b_1 + b_0 + a_1}$$



(a) out-degree distribution.



(b) in-degree distribution.



(c) bi-degree distribution.

Figure 6.1: The degree distributions of Twitter snapshot.

6.6 A new Preferential Attachment Model for Twitter

We present here a new preferential attachment model for Twitter, whose analysis led us to the introduction of our framework. The model takes into account some properties exhibited by the exploration of a Twitter snapshot of 505 million nodes and 23 billion edges, which we first report in Section 6.6.1. The model in then given in Section 6.6.2.

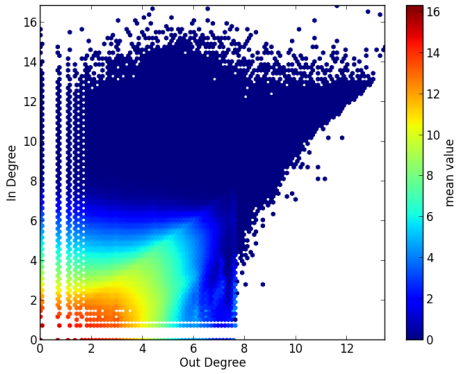
6.6.1 Experiments on Twitter

We report on experiments performed on the degree distributions of the Twitter social graph.

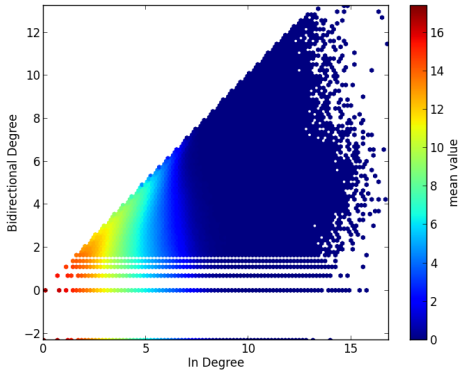
Dataset. To analyze the graph connecting Twitter's users, we used the Twitter snapshot described in [14] and made available by the authors. The authors successfully crawled all the Twitter accounts (except 6% of protected accounts) in 2012 and collected their followed users. The social graph built has 505 million accounts connected with 23 billion arcs. The authors then unveiled the macroscopic structure of the graph.

In the following, we study the degree distributions of the graph. We consider here nodes in Twitter's Largest Strongly Connected (LSC) component. Indeed, it involves around half of the total users, more than 96% of the following and follower links, and 98.05% of the tweets are for accounts in the LSC [14].

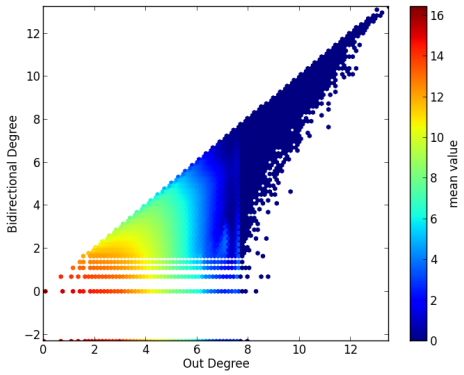
Strong presence of bi-directional links. We found that 35% of the arcs are involved in bi-directional links. Furthermore, 71% of the vertices have at least one



(a) out-degree versus in-degree.



(b) in-degree versus bi-degree.



(c) out-degree versus bi-degree.

Figure 6.2: Heat-map of the correlations between the degree distributions in Twitter snapshot.

bi-directional link. Moreover, these bi-directional links have an important social meaning (e.g., friendships, ...). An adequate model for Twitter should thus take into account bi-directional links.

Relations between node degrees: Power law and Strong Degree Correlation. We studied further the relation between node degrees. It is well known in a undirected social graph that a node with high degree has more chance to attract new connections than a node with low degree. Is this also the case for in-, out-, and bi-directional- degrees? As expected, we found that it is the case and we exhibit a heavy-tail phenomena for the three of them as shown in Figure 6.1.

But more importantly and newer, does a node with high in-, out-, or bi-directional-degree have more chance to attract arcs of the other kinds? We analyze the correlations between the different degrees. To give the reader a perception of them, we plot heat-map plots in Figure 6.2. In each plot, the considered degree distributions are on the axes and the colours are used to indicate the number of vertices, e.g., dark blue means few vertices. We also quantify the correlations by computing the Pearson coefficients between the three degree distributions.

We find a very low correlation between in-degree and out-degree (see Figure 6.2a). Indeed, their Pearson coefficient is around 0.15. This reflects the fact that content publishers (e.g., a celebrity) may have a lot of followers without following themselves a lot of people. Symmetrically, this can be explained by the social fact that choosing

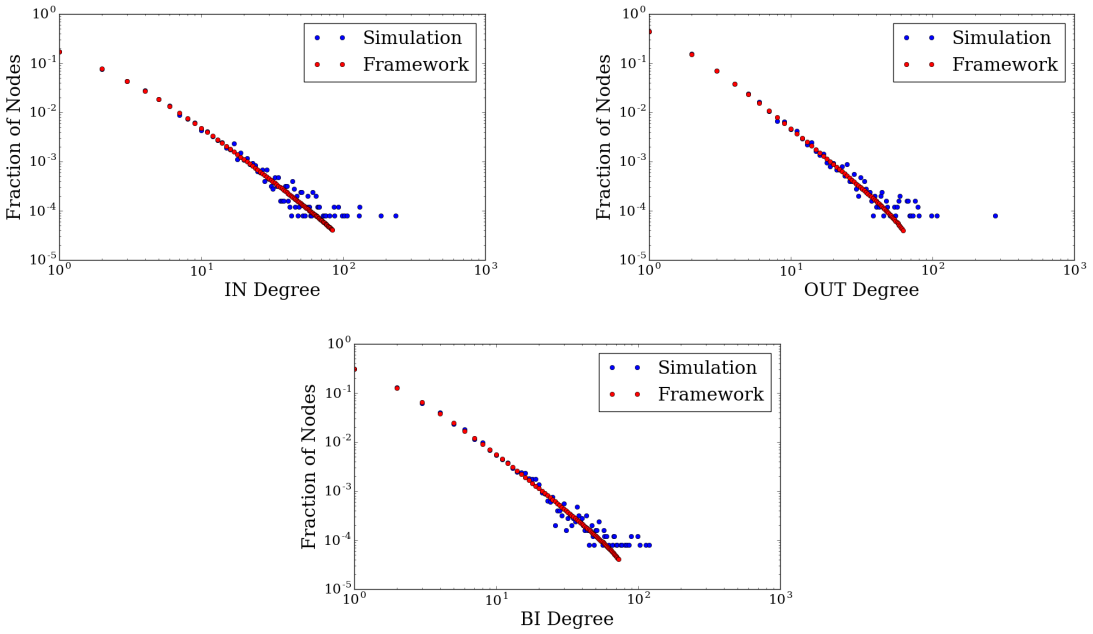


Figure 6.3: Degree distributions of the Twitter model. Comparison between simulations with 1 M nodes and the framework with a truncation box size of 100. Parameters: $\alpha = 0.25$. $\beta = 0.35$. $\delta_{in} = 1$. $\delta_{out} = 1$.

to follow a large number of users does not make us more interesting. Note that this validates random graph models such as the ones of [9] in which independence between in- and out- degrees are supposed for the computations.

However, the relation between out-degree and bi-degree is very different and interesting. We find a Pearson coefficient around 0.95, showing a very strong correlation. This can also be observed in Figure 6.2c. Note that we expected such a correlation. Indeed, there is a tendency in Twitter for a category of users to follow back the people who follow them in order to gain followers, hoping that other users will do the same [14]. But, we did not expect such a strong correlation.

For the last relation between in-degree and bi-degree, we find a very low correlation (see Figure 6.2b). The value of the Pearson coefficient is 0.15. This can be explained by the fact that very popular users do not follow back (they do not need to do it) and that content popularity is different from friendship popularity (bi-directional) on a social network.

Note that if we had found no correlation between out-degree and bi-degree, the degree distribution of Twitter could have been taken into account with a very simple model: take a random directed graph according e.g., to [9] and a random undirected graph from [6], and merge them together. However, the strong correlation found forces us to propose new models.

We are thus in need of a new model taking into account the large number of bi-directional links and the degree correlation.

6.6.2 Random directed graph models

We propose a new model of preferential attachment to explain the large presence of bi-directional links in directed Online Social Networks like Twitter.

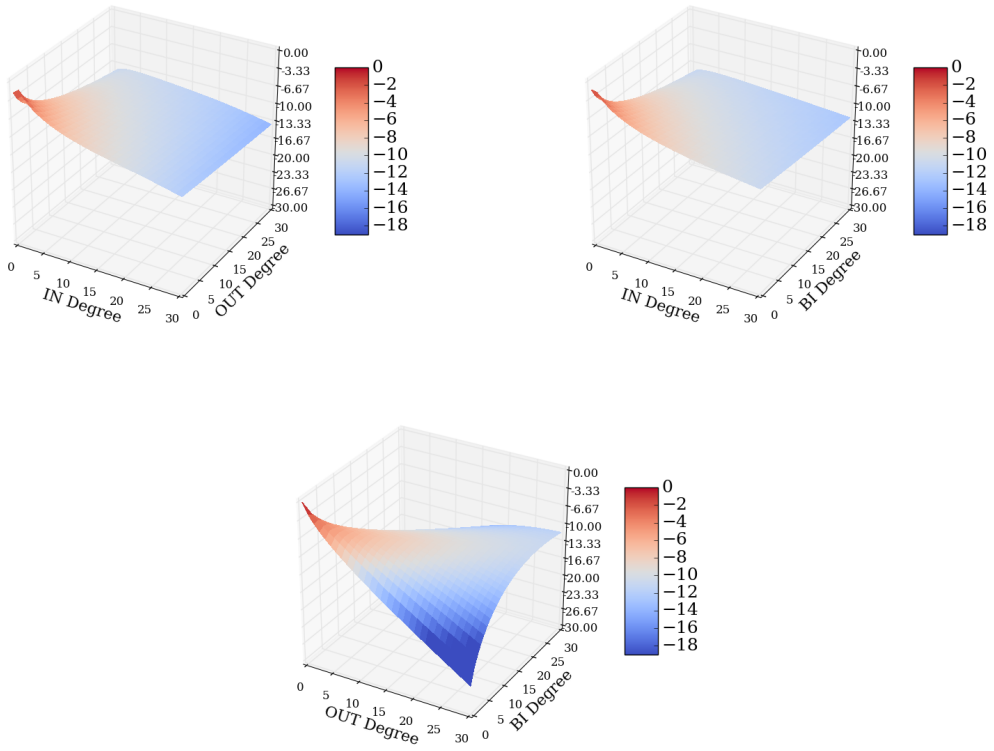


Figure 6.4: Correlations between in-, out-, and bi-degrees in the model with correlated bidirectional links.

We start at the initial time t_0 from an initial directed graph G_{t_0}. The graph then grows with two kinds of events: an *arc event*, during which a (single) arc or a bi-directional (double) arc is added between two existing vertices; and a *vertex event*, during which a new vertex is added, as well as a single or double arc linking it with an existing vertex. Two important features of the model are (i) that we choose the two ends of a bi-directional arc according to their out-degree and (ii) that we do not consider bi-directional arcs when computing the in-degree of a node. The reason is that we found a strong correlation between bi-degree and out-degree in Twitter, and no correlation with the in-degree.

Model Description

We consider a model that starts with a directed graph which grows by adding single or double arcs at each time step. Furthermore, at each step a vertex may or may not be added. A vertex event happens with probability α and an edge event happens with probability $1-\alpha$. When a link appears, it is a double arc with probability γ and a single arc with probability $1-\gamma$.

Let t_0 be the initial time, with $t_0 \geq 1$. We denote by $G(t)$ the graph at time $t \geq t_0$, $e(t)$ its random number of arcs and $n(t)$ its random number of vertices. We define $\overline{d}_{in}(v)$ as the number of incoming arcs at a node v that are not involved in a bi-directional relation (when the total in-degree $d_{in}(v)$ includes them). In what follows, the *choice* of a vertex v of G_t by either $\overline{d}_{in}+\delta_{in}$ or $d_{out}+\delta_{out}$ means to choose the vertex v with some probability proportional to either $\overline{d}_{in}+\delta_{in}$ or $d_{out}+\delta_{out}$.

Specifically, $Pr[v = v_o] = \frac{d_{out}(v_o)+\delta_{out}}{e(t)+\delta_{out}n(t)}$ is according to out-degree, where $d_{out}(v_o)$ indicates the out-degree of the vertex v_o in the graph $G(t)$. For the in-degree, the probabilities are similar by interchanging d_{out} with $\overline{d}_{in}$ and δ_{out} with δ_{in}.
The graph $G(t)$ evolves in $G(t+1)$, for $t \geq t_0$, according to the following rules:

- (A) With probability $\alpha(1-\gamma)$, add a new vertex v together with an arc from v to an existing vertex w, where w is chosen according to $\overline{d}_{in} + \delta_{in}$;
- (B) With probability $\alpha\gamma$, add a new vertex v together with an arc from v to an existing vertex w and one arc in the inverse direction, where w is chosen according to $d_{out} + \delta_{out}$;
- (C) With probability $(1-\alpha)(1-\gamma)$, add a new arc from an existing vertex v to an existing vertex w, where v and w are chosen independently, v according to $d_{out} + \delta_{out}$ and w according to $\overline{d}_{in} + \delta_{in}$;
- (D) With probability $(1-\alpha)\gamma$, add two arcs between existing vertices v and w, where v and w are chosen independently and according to $d_{out} + \delta_{out}$;

We define $\alpha, \gamma, \delta_{in}$, and δ_{out} to be non-negative real numbers, with α, γ in the range $[0, 1]$. Studies in [14] reveal the existence of a large number of vertices with out-degree or in-degree equal to 0 and which are connected to other vertices. We thus need that they can be chosen as the end of an arc with a positive probability during the preferential attachment process. To this end, we borrow from the model in [9] two technical constants, δ_{in} and δ_{out}. For future work, we aim at varying the value of δ_{in} and δ_{out} for each vertex depending on its component: e.g., a vertex in the disconnected component may have $\delta_{in}, \delta_{out} = 0$.

Now we identify the random variable $x_{i,j,k}(t)$ as the number of vertices in $G(t)$ with:

- $i + k$ in-degree, where i indicates the number of incoming arcs that are not involved in a bi-directional relation.
- $j + k$ out-degree, where j indicates the number of outgoing arcs that are not involved in a bi-directional relation.
- k indicates the number of arc pairs involved in bi-directional relations.

For sake of simplicity of the computations, we allow multiple arcs, loops and account for bidirectional arcs that come out only from the events (B) and (C). This is a classical way to carry out the analysis and this does not affect the results.

Computation of the model distributions: a case of three-dimensional state.

We compute here the degree distributions in the random Twitter model. In this situation, the stationary distribution satisfies the following recurrence equation:

$$\begin{aligned}
&c_1(i-1+\delta_{in})(S_{i,j,k} - S_{i-1,j,k}) \\
+&c_2(j+k-1+\delta_{out})(S_{i,j,k} - S_{i,j-1,k}) \\
+&c_3(j+k-1+\delta_{out})(S_{i,j,k} - S_{i,j,k-1}) \\
+&(c_1+c_2+c_3)S_{i,j,k} \\
=&\ 0
\end{aligned}$$

with $c_1 = \frac{1-\gamma}{1+\gamma+\delta_{in}\alpha}$, $c_2 = \frac{(1-\gamma)(1-\alpha)}{1+\gamma+\delta_{out}\alpha}$, $c_3 = \frac{\gamma(2-\alpha)}{1+\gamma+\delta_{out}\alpha}$.

Observe that if we were only considering the two dimensions i and $j+k$, then we would go back to the previous two-dimensional state case, with $a_0 = c_1$, $b_0 = c_1 \cdot \delta_{in}$, $a_1 = c_2 + c_3$, $b_1 = (c_2 + c_3) \cdot \delta_{out}$. Let $\hat{S}_{i,j+k}$ be the corresponding stationary distribution. Then,

$$S_{i,j,k} = \hat{S}_{i,j+k} \cdot \mathbb{P}r[j \mid j+k].$$

The latter corrective term corresponds to the probability of j successful events among $k+j$ Bernouilli trials with $c_2/(c_2+c_3)$ success probability. Hence,

$$S_{i,j,k} = \binom{j+k}{j} \cdot \frac{c_2^j c_3^k}{(c_2+c_3)^{j+k}} \cdot \hat{S}_{i,j+k}.$$

6.6.3 Validation

In this Section, we validate the fast convergence of the framework with truncation using simulations. We implement two different preferential attachment models: the one presented in Section 6.6.2, with correlated bi-directional links, and a model of graphs with edge removals, described below. We then compare the degree distributions given by the framework (stationary distribution of a truncated infinite Markov chain) with the average degree distributions over sets of random graphs.

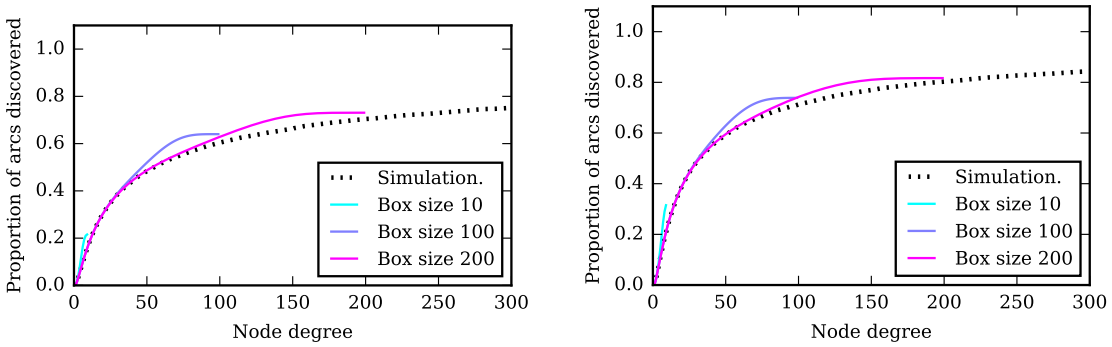
One of the advantages of the framework is that we very quickly obtain the degree distribution for any size of graphs, when the simulations of graphs with millions of vertices and billions of edges are taking prohibitive time.

Simulation settings. The simulation consists in building a small number (20) of random graphs with 1 million nodes. We then average the degree distributions over the built random graphs.

Computing the degree distributions given by the framework. To compute the stationary distribution of the truncated chain, several methods can be used. We tested two methods: (i) inverting the matrix using a linear solver; (ii) using an iterative computation. Method (ii) is very slow. We thus apply Method (i) to produce the results presented in the following. We used IBM CPLEX solver [17]. To show the efficiency of the truncation, we tested different sizes for the box used to truncate the infinite Markov chain, see Figure 6.5.

Models with correlated bi-directional edges.
Validation of the framework. We first validate the framework by comparing the distributions of in-degree, out-degree, and bi-degree given by the framework with the ones obtained from simulation. We see in Figure 6.3 that framework and simulations almost perfectly match.

Correlation. In Figure 6.4, we show the joint distributions of in- and out-degrees, of in- and bi-degrees, and of out- and bi-degrees given by the framework. We observe the strong correlation in the model between out and bi-degrees, as pointed out by the "red" diagonal in the Right plot. This is desired, as this strong correlation is observed in Twitter. This is due to the choices of the end-vertices of bidirectional links according to the out-degree in the model. On the contrary, in-degree and bi-directional degree (and in-degree and out-degree) are a lot less correlated (Left and Middle plots). This is also desired as observed on Twitter, and this is due to the selection of the head of a directed simple link according to the in-degree of vertices



(a) in-degree

(b) out-degree

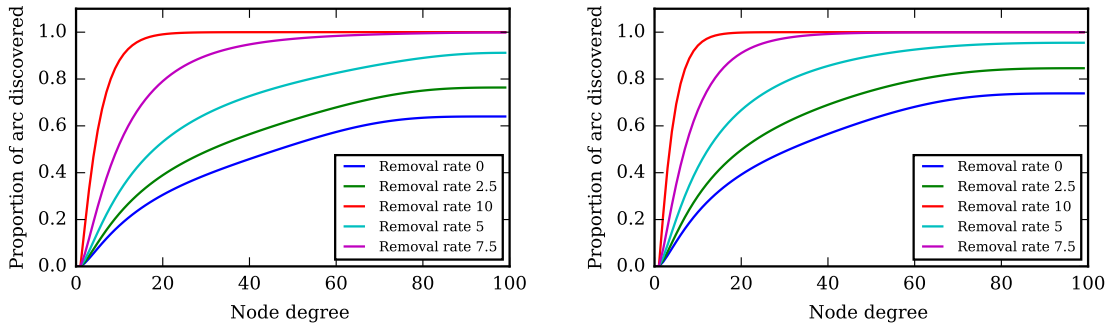
Figure 6.5: Study of the truncation box size of the infinite Markov chain: cumulative in-degree and out-degree distributions.

(excluding bidirectional links). We still observe a correlation between in and bi-directional degree. This is due to the impact of the date of apparition of vertices. Indeed, a vertex arriving at the beginning of the preferential attachment process has more chances to have high in-, out-, and bi-degrees.

Models with edge removal. We consider a generalization of the two-dimensional state model of Bollobás et al. [9], in which edges can be added, but also *removed*, during time. In an online social graph, this would correspond to the loss of interest from a user to the content of another user. We introduce in the model a probability r, which is the probability for each edge to be removed at each time step. Note that, for this model, there exists no simple recurrence equation to compute the degree distributions of the preferential attachment model. We show in the following that we can easily evaluate numerically the distributions using our framework.

Validation of the truncation process. We tested the precision of the truncation. We study the cumulative proportion of arcs discovered considering nodes of in-degree or out-degree lower than a given value. Results for three sizes of truncation boxes are shown in Figure 6.5. We observe that the truncation works very well for degrees up to around one third of the box size. As expected, values for degrees closer to the box size are not very precise. A user of the framework should thus choose a box size larger than the maximum degree she wants to observe. A precise study of the adequate box size is left for future work.

Study of the impact of edge removal. We discuss now the impact of edge removal on the degree distributions. We compare in Figure 6.6 the cumulative distribution functions for different rates of edge removal, for in-degree in Figure 6.6a and out-degree in Figure 6.6b. We see that the introduction of edge removals concentrates the degree distribution towards low degrees. For a removal rate equal to 10 (and also equal to the edge arrival rate), almost all nodes have an in-degree smaller than 20. With a removal rate of 0, almost 40% of nodes have an in-degree greater than 100. The effect is similar on out-degrees.



(a) in-degree (b) out-degree

Figure 6.6: Study of the model with edge removals. Cumulative in-degree and out-degree distributions for different removal rates.

6.7 Conclusion and open perspectives

In this Chapter, we studied some properties of one of the biggest directed snapshots of social networks available nowadays, a graph of the social interactions of Twitter, made available to us by the authors of [14]. We have noticed that the graph has a high number of bidirectional links (around 35% of the edges) and that, if the in-degree is uncorrelated with the out- and bi-degrees, the out-degree and bi-degree are strongly correlated.

It led us to propose a new preferential attachment model to take into account these properties of Twitter. However, classical ways to analyze preferential attachment were not powerful enough to compute the degree distributions of our model. We thus proposed a new theoretical framework to compute the degree distributions of a broad set of preferential attachment models. The key idea is to reduce the computations to the analysis of the stationary distribution of a continuous Markov process. It made it possible for us to derive both the independent or joint distributions (e.g., the probability to have a given in-degree and out-degree at the same time), for almost any model with node events and edge events (including those with edge removals). As an example, we used this framework to compute, for the first time in literature to our knowledge, the combined degree distribution of the Bollobás & al. model [9], as well as our Twitter model with three correlated dimensions. We do believe this framework can be used on an even larger set of preferential attachment models. In particular, we think it would be possible to push the study of the analytical solutions of preferential attachment models to states of dimensions $\mathbb{N}^k$, with similar computations as done in this Chapter.

Bibliography

[1] SR Adke. 201. note: a multi-dimensional birth and death process. *Biometrics*, 20(1):212–216, 1964.

[2] William Aiello, Fan Chung, and Linyuan Lu. A random graph model for power law graphs. *Experimental Mathematics*, 10(1):53–66, 2001.

[3] Mohamed Akkouchi. On the convolution of exponential distributions. *J. Chungcheong Math. Soc*, 21(4):501–510, 2008.

[4] Konstantin Avrachenkov, Alexey Piunovskiy, and Yi Zhang. Markov processes with restart. *Journal of Applied Probability*, 50(4):960–968, 2013.

[5] Lars Backstrom, Paolo Boldi, Marco Rosa, Johan Ugander, and Sebastiano Vigna. Four degrees of separation. In *Proceedings of the 4th Annual ACM Web Science Conference*, pages 33–42. ACM, 2012.

[6] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[7] Albert-László Barabási, Réka Albert, and Hawoong Jeong. Mean-field theory for scale-free random networks. *Physica A: Statistical Mechanics and its Applications*, 272(1):173–187, 1999.

[8] Albert-László Barabási, Réka Albert, and Hawoong Jeong. Scale-free characteristics of random networks: the topology of the world-wide web. *Physica A: statistical mechanics and its applications*, 281(1-4):69–77, 2000.

[9] Béla Bollobás, Christian Borgs, Jennifer Chayes, and Oliver Riordan. Directed scale-free graphs. In *Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 132–139. Society for Industrial and Applied Mathematics, 2003.

[10] F. Chung and L. Lu. *Complex graphs and networks*. AMS, 2006.

[11] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

[12] Colin Cooper and Alan Frieze. A general model of web graphs. *Random Structures & Algorithms*, 22(3):311–335, 2003.

[13] Maksym Gabielkov and Arnaud Legout. The complete picture of the twitter social graph. In *Proceedings of the 2012 ACM conference on CoNEXT student workshop*, pages 19–20. ACM, 2012.

[14] Maksym Gabielkov, Ashwin Rao, and Arnaud Legout. Studying social networks at scale: macroscopic anatomy of the twitter social graph. In *ACM SIGMETRICS Performance Evaluation Review*, volume 42, pages 277–288. ACM, 2014.

[15] Xianping Guo and Onésimo Hernández-Lerma. Continuous-time markov decision processes. In *Continuous-Time Markov Decision Processes*, pages 9–18. Springer, 2009.

[16] Oskar Hagberg and Carsten Wiuf. Convergence properties of the degree distribution of some growing network models. *Bulletin of Mathematical Biology*, 68(6):1275, 2006.

[17] IBM. Ibm cplex solver https://www.ibm.com/analytics/cplex-optimizer.

[18] Michael Knudsen and Carsten Wiuf. A markov chain approach to randomly grown graphs. *Journal of Applied Mathematics*, Vol. 2008, 2008.

[19] Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, D Sivakumar, Andrew Tomkins, and Eli Upfal. Stochastic models for the web graph. In *Foundations of Computer Science, 2000. Proceedings. 41st Annual Symposium on*, pages 57–65. IEEE, 2000.

[20] Jure Leskovec and Eric Horvitz. Planetary-scale views on a large instant-messaging network. In *Proceedings of the 17th international conference on World Wide Web*, pages 915–924. ACM, 2008.

[21] Colin McDiarmid. On the method of bounded differences. *Surveys in combinatorics*, 141(1):148–188, 1989.

[22] Seth A Myers, Aneesh Sharma, Pankaj Gupta, and Jimmy Lin. Information network or social network?: the structure of the twitter follow graph. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 493–498. ACM, 2014.

[23] Mark EJ Newman. Random graphs as models of networks. *arXiv preprint cond-mat/0202208*, 2002.

[24] D Shi, Liming Liu, SX Zhu, and H Zhou. Degree distributions of evolving networks. *EPL (Europhysics Letters)*, 76(4):731, 2006.

[25] Dinghua Shi, Qinghua Chen, and Liming Liu. Markov chain-based numerical method for degree distributions of growing networks. *Physical review E*, 71(3):036140, 2005.

[26] Flora M Spieksma. Countable state Markov processes: non-explosiveness and moment function. *Probability in the Engineering and Informational Sciences*, 29(4):623–637, 2015.

[27] Flora Margaretha Spieksma. Kolmogorov forward equation and explosiveness in countable state Markov processes. *Annals of Operations Research*, 241(1-2):3–22, 2016.

[28] Tian Wang, Hamid Krim, and Yannis Viniotis. A generalized Markov graph model: Application to social network analysis. *IEEE Journal of Selected Topics in Signal Processing*, 7(2):318–332, 2013.

[29] Nicholas C Wormald. The differential equation method for random graph processes and greedy algorithms. *Lectures on approximation and randomized algorithms*, pages 73–155, 1999.

[30] George Udny Yule. Ii.—a mathematical theory of evolution, based on the conclusions of dr. jc willis, fr s. *Philosophical transactions of the Royal Society of London. Series B, containing papers of a biological character*, 213(402-410):21–87, 1925.

Chapter 7

Conclusion

We have discuss through those hundred and fifty pages various works which all follow the same Ariadne's string, namely the will to study real-world networks with complex properties and to model them as accurately as possible. I present here a (brief and non exhaustive) summary of the main properties highlighted by the real-world network studies:

- **Twitter network:** The main results can be found in Sections 3.4, 3.3.1 and 6.6.1. This directed network exhibits:
  - A **high presence of bidirectional links**, with around 32% of the edges implied in a bidirectional link;
  - **In- and bi-degree distributions following power-law distributions**. The same holds for the **out degrees** in the tail of the distribution, while lower degrees have an atypical shape - in particular with a huge spike around 2000, due to Twitter's policies.
  - A **high correlation between in- and out-degrees and between out- and bi-degrees**, but a **low correlation between in- and bi-degrees**;
  - A **high value of the interest clustering coefficient**, both in the whole graph and the in graph without bidirectional link; **other classical directed clustering coefficients are high in the whole graph and in the mutual graph but low in the graph without bidirectional edges**. This confirms the idea that Twitter is both a social network (due to its high triangle clustering in bidirectional graph) and a network of information (due to its high interest clustering coefficient in the graph without bidirectional links).
- **Copublication network:** The main results can be found in Sections 5.5, 5.6 and 5.7.2. This network exhibits:
  - **One large connected component**, containing 95% of the nodes and 99% of the hyperedges;
  - **47 communities** of size bigger than 100 (estimated with the Leiden algorithm [7]);
  - A **high modularity** both for the flatten graph (around 0.80) and the hypergraph (around 0.63);

  - A **degree distribution following a power-law** (in first approximation);
  - A **size of communities distribution apparently following a power-law** too (see Figure 5.1). Note that, if I did not use this property in this thesis, this assumption is used in various models such as the LFR benchmark [4].

I also summarize the presented models in order to give an idea of which networks are covered by the results of this thesis. The presented models enable the construction of:

- graphs with any wanted degree distribution (Section 4.3);
- directed graphs with a power-law degree distribution and a high value of interest clustering coefficient (Section 3.6);
- directed graphs with in- and out- power-law degree distributions, high proportion of bidirectional edges and correlations closed to the ones of Twitter (Section 6.6);
- hypergraphs with a power-law degree distribution and the presence of communities (Section 5.4).

If the results presented in this thesis are interesting in themselves, they also open many perspectives which I hope to explore in future works:

Merge of the models: First of all, each presented model covers different types of networks in order to give a panel of possibilities. In the future I would be interested in merging the proposed models into a really general model encompassing a lot of wanted properties. In particular, adding a general attachment function as proposed in Chapter 4 to the model for hypergraph with communities presented in Chapter 5 would be really interesting in order to be able to build hypergraphs with communities and general degree distributions. I expect the proof for the degree distribution of the model without communities (Section 5.3.2) to be similar, but with a condition on the chosen probability distribution as the one we have in Theorem 2. However, in a similar way to what is presented in Section 5.6, the degree distribution of the model with communities will be the sum of all communities distributions. This sum might give unexpected results, which further work could try to deal with.

Apply the introduced tools: A few novel tools has been introduced in this thesis, which can be applied in future works. In particular, the interest clustering coefficient could be applied on other large directed networks with interest links, e.g. bigger datasets than the ones used in Section 3.5 for Instagram and scientific citations networks, but also email, phone calls or text message networks, etc. The framework presented in Chapter 6 might be used to compute the degree distributions of other complex models, in particular joint degree distributions of models with more than 1 dimension. For instance we are currently using it to compute the degree distribution of a general model merging directed, undirected and bi-directional edges.

I also believe the relation between the degree distribution and the attachment function expressed in Equation 4.1 from Theorem 2 opens a lot of perspective. I remind Theorem 2 here:

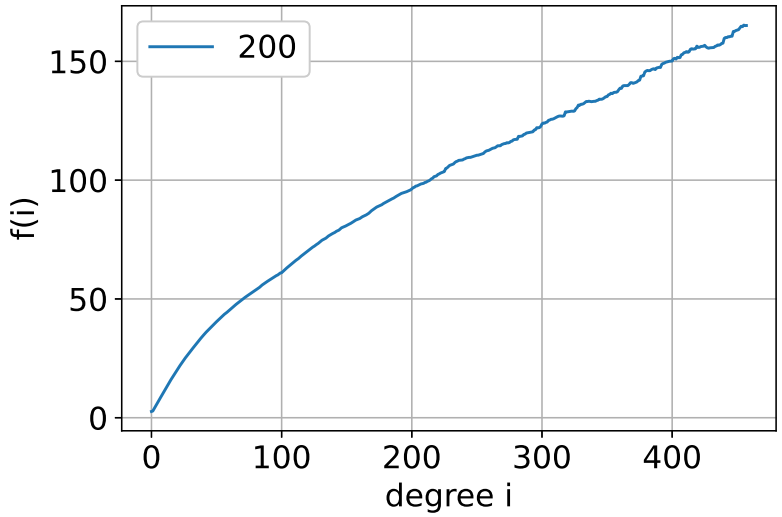


Figure 7.1: Attachment function obtained from Equation 4.1 for the copublication hypergraph degree distribution. Each point is a mean on the 200 points around it.

Theorem 2. *Let P be a probability distribution of mean μ and such that the function $h(i) = \frac{P(k>i+1)}{P(i+1)} - \frac{P(k>i)}{P(i)}$ is bounded. In the model proposed in Section 4.3.1, if $p = \frac{1}{\mu}$ and the attachment function is chosen as:*

$$\forall i \geq 1, f(i) = \frac{P(k > i)}{P(i)}, \tag{4.1}$$

then the degree distribution of the created graph is distributed according to P.

To the best of my knowledge, such a general link between those essential components of networks have never been expressed in such a general way, and the simplicity of the relation is really promising for the emergence of concrete, simple connections. I even hope that this might bring some knowledge on how real-world networks emerge and under which rules they evolve. For instance, Figure 7.1 presents the computed attachment function using Equation 4.1 for the copublication hypergraph degree distribution. In order to get rid of the noise, I did a moving average of size 200. We see that the attachment function indeed seems linear at the end - thus leading to a power-law degree distribution - but that the beginning is not linear. Some deeper studies could be conducted using those insights. This gives possible insights on how nodes evolve in the network, supposing the evolution follows the model presented in Section 4.3. Remind that Equation 4.1 only holds for the model proposed in Section 4.3. However, I believe some equivalent equations can be expressed for other random growth models.

Deeper study of the real-world networks: The two datasets we have at our disposal are really rich in information, and further studies might highlight other properties. Among them, studying communities in the Twitter's network would be really instructive. This network being directed, the definition of a community is not straight-forward. Indeed, a link between two nodes does not mean that they are close to each other, but that one of them is interested in the other. Two nodes will rather belong in the same community if they share multiple common interests, i.e., if they have similar out-neighborhood. Studying communities thus go through a

previous transformation of the network - see [5] for a survey of the study of directed communities. In Twitter, we expect to have strong clusters due to languages, and other sub-clusters inside them representing different interests - music, football, ... Verifying those assumptions is an open future work.

Finally, the copublication network is ideal to study the impact of funding in research. Initially, the Scopus publications dataset has been crawled under the SNIF project, implying the Inria and I3S laboratory, the SKEMA Buisness School, and the GREDEG research laboratory of economy, with the aim of studying the impact of funding on productivity and multidisciplinary. The copublication graph can be used to follow researchers that received a grant at some given time, and see how their publications have changed from this date. Those evolutions can be compared with researchers who have same publication behaviors but did not received a grant, in order to compare the difference of evolution between both. The comparison can be made using a metric focusing on the property we want to study: publication rate, pluridisciplinarity publications, ... Some preliminary studies seem to show that researchers who get a grant tend to do more collaborations with authors they have never published before, in comparison to not-granted researchers with same publication rates.

To complete this study, the model proposed in Section 5.4 can be improved in order to introduce funding. We plan to add fitness to nodes, similarly to the model proposed by Bianconi and Barabási in 2001 [1, 2]. Each node would have a probability to receive a new edge proportional to its fitness. The idea is to change the fitness of some nodes taken randomly after some time, in order to model the arrival of grants according to those nodes. Then, we study the degree evolution of those nodes in comparison to some nodes for which fitness didn't changed - i.e., who did not received grants.

Finally, we would be interested to use the copublication hypergraph to quantify inequalities between men and women in science, and the impact of gender equality's policy on publications. In recent years, politic rules have been applied in order to decrease the observed inequalities between men and women in almost all domains. This also applied for research, with women's only funding, equal proportions to some board of examiners, ... The copublication network seems ideal to study the impact those funding had on publications and collaborations, and if we observe a difference since this gender equality's policy has been developed.

Study of properties for directed graphs and hypergraphs: On a longer-term vision, dealing with directed graphs and hypergraphs is an open-field in which a lot of work remains to be done. If most of the developed tools are built for undirected graphs, a generalization to other types of graphs often stays an open problem. This is not an easy task: some problems for which we know simple polynomial algorithm for undirected graphs become NP-complete for directed graphs. It is the case for instance for the *k-linkage problem* which, for a graph G and k pairs of vertices of G, decides if there are k mutually vertex-disjoint paths of G joining the pairs [6, 3]. In some cases, the generalization of the metrics is also not straight-forward, as discussed above for community detection. This is thus an active and important direction of research. As a contribution to this field, Chapter 3 and Chapter 5 propose models for hypergraphs and directed graphs. Chapter 3 focuses on directed clustering coefficients by comparing different definitions of it. I also started to compute the different definitions of directed clustering coefficients in random directed models such as the

directed configuration model and the directed erased configuration model (configuration model with deletion of loops and multiedges). The computation is based on the one conducted by van der Hofstad et al. [8], but has to deal with complications coming from the directed aspect of the model (there is a correlation between in-in degrees while not between in-out degrees). This is a work in progress in collaboration with Guillaume Ducoffe, Frédéric Giroire, Stéphane Pérennes and Małgorzata Sulkowska.

Bibliography

[1] Ginestra Bianconi and A-L Barabási. Competition and multiscaling in evolving networks. *EPL (Europhysics Letters)*, 54(4):436, 2001.

[2] Ginestra Bianconi and Albert-László Barabási. Bose-einstein condensation in complex networks. *Physical review letters*, 86(24):5632, 2001.

[3] Steven Fortune, John Hopcroft, and James Wyllie. The directed subgraph homeomorphism problem. *Theoretical Computer Science*, 10(2):111–121, 1980.

[4] Andrea Lancichinetti, Santo Fortunato, and Filippo Radicchi. Benchmark graphs for testing community detection algorithms. *Physical review E*, 78(4):046110, 2008.

[5] Fragkiskos D Malliaros and Michalis Vazirgiannis. Clustering and community detection in directed networks: A survey. *Physics reports*, 533(4):95–142, 2013.

[6] Neil Robertson and Paul D Seymour. Graph minors. xiii. the disjoint paths problem. *Journal of combinatorial theory, Series B*, 63(1):65–110, 1995.

[7] V.A. Traag, L. Waltman, and N.J van Eck. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep. UK*, 9(5233), 2019.

[8] Remco van der Hofstad, Pim Van der Hoorn, Nelly Litvak, and Clara Stegehuis. Limit theorems for assortativity and clustering in null models for scale-free networks. *arXiv preprint arXiv:1712.08097*, 2017.