Power-law phenomena in Bayesian nonparametrics

Francesca Panero
St. Peter's College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy
Hilary 2022

Statement of Originality

I hereby declare that except where specific reference is made to the work of others, the contents of Chapters 1, 2, 3, 4 and 6 are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university.
Part of chapter 5 has been submitted by me in July 2017 as thesis for the degree of MSc in Stochastics and Data Science at Università degli Studi di Torino[1]. The work of chapter 3 of the MSc thesis became part of appendix A.2, A.3, A.4 and A.5 of the paper presented in chapter 5 of this thesis.
My personal contributions are as outlined in the authorship forms at the end of each chapter. This dissertation is my own work except as specified in the text, acknowledgements, forms and papers.

Francesca Panero
Hilary 2022

[1] A copy of the MSc thesis can be found at the following link: `https://github.com/francescapanero/MSc_thesis.git`

A mia nonna, Giò,
e a tutte le donne
che non hanno avuto e non hanno la possibilità di studiare.

To my grandma, Giò,
and to all the women
who did not and do not have the opportunity to study.

Acknowledgements

"It takes a village to raise a child"
African proverb

And to do many other things, I would add.

On the 14th of October 2017 I was running (no surprise, I was late) towards the Sheldonian Theatre to matriculate, wondering if I would have been able to make it to the end of the DPhil. Now that I am close to it, I can acknowledge formally each and every one who has been close to me during these years. A PhD often feels like a life-changing yet - especially during a pandemic - incredibly lonely trip. So, sometimes, you really need your village.

First of all, my deepest gratitude goes to François and Judith, my supervisors. Week after week, they took me with them in this research trip. I bothered them with countless questions, and they always helped me through. As they were pushing me to find and break the next bug in the code, so they did with my limits. They advised me on work choices, with a special mention to Judith for being close when the stress for work decisions skyrocketed. They have been buddies of pints and lunches in the graveyard. Looking at them discussing research ideas will be one of the best memories of this PhD.

I was part of the last cohorts of European students lucky enough to benefit from the pre-Brexit access to public funds to cover their PhD studies. I thank EPSRC for the scholarship that supported me during these years.

An important recognition goes to Prof. Favaro, who has been much more than a MSc supervisor and has always captivated me with his projects (research or DIY). It seems you cannot get rid of me easily! But he is not the only one I wish to continue my research exploration with. Thank you to Prof. Dunson

for our collaboration at Duke that, despite being cut short by the pandemic, opened my eyes to interesting applied research directions. To Federico, my long-standing friend and research crush, who keeps on running away but whom I will eventually catch and publish something with. Thank you to Tommaso, friend in Duke and Milan, from whom I have learned a lot and I hope to keep on doing so. To Marco Scutari, who offered me an internship at IDSIA, allowed me to work on one of the topics that interest me the most and supported me in the search of a job. Thank you to Prof. Lambiotte, who assessed me during the transfer and confirmation of status, providing insightful comments and points of view.

My PhD trip would have not be the same without my OxWaSP friends. Special mention goes to James, who bravely came to Italy without a phone and is the improptu friend who I'd like to have next in many concerts. To Marco, Francesca and Suzie, my (almost) Italian support group and bridge buddies, who abandoned me to move to Coventry but have been willing to spend 15 days in quarantine (and catch COVID-19, nevertheless) to come to my wedding. To Kaspar, my wise college dad and R Users Group friend. To Xenia: friends, wedding twins, mutual cheerleaders and soon colleagues. To Déborah, my BNP, ED&I and St. Peter's buddy...a sister in Oxford.

The Department of Statistics has been theatre of numerous adventures with great people. I have learned immensely from all the chats, seminars and classes I participated to. Thanks to all the staff, who made the department more than a workplace. To Mariagrazia, for taking on the RUG and sharing tea and pizza with me. To Joanna, the best possible programme administrator, with whom I share a love for chats, singing in choirs and marrying musicians. Thank you to Beverley, tireless organiser of the activities of the department and the wonderful Mental Health Awareness week, for helping out with ED&I activities and being an enthusiastic member of the choir and green team. To all the members of the StatsChoir and the EquiStats committee: it has been fun and I learned so much from all of you.

I have lived in Oxford for less time than I wanted, but my life has nevertheless crossed, let go and tangled many threads. A lot of them were part of St. Peter's College and its boat club, where I lived incredibly enriching and fun (and physically exhausting) experiences. To Alessa and Nils, with whom I share the joy and doubts of expat couples. To Sofia and Nick: it is impossible to know where you will be the next month, and this is why I will not lose you. To Narain, who is a friend and has hosted me with Liliana when I did not have a house in Oxford.
If my accent is still terrible, you can also blame my Little Italy. To Giuseppe and Jonida, whose friendship extends in time and space: I could not ask for more. Matteo and Virginia, whom I miss dearly and I cannot wait to hug again. To Liliana, my other half in St. Clement's, that has shared with me many Oxford experiences and with whom many life decisions have been discussed. To Marion and Andrea, whose support has been invaluable and are the reason why a part of my heart will always be in London (before I visit the baita, of course). To Alberto and Marco, a mixture of Bergamo, Moncalieri, London and Rome that never ceases to excite me.

Some bonds have not been defeated by borders. To Elisabetta, Francesca, Roberta and Sofia, who have been there for 15 years and will stick around for much more. We grow old, but in my mind we are still eating salad from a bag, lying down under the Eiffel Tower. To Alessandra, Giulia and Marta, without whom I am not sure I would have gone so far into the Maths world. To Eleonora, the memory I do not have, the friend I could call every day of my life: I am so looking forward to our future. To Alice, Luca and Sarah, the only ones who can make me appreciate a virtual office, colleagues in PhD (mis)adventures, home away from home. To Iacopo, the not-anymore +1. To Cecilia, who became much more than a Stats friend. To Chiara, an incredibly interesting woman I'm happy to have the opportunity to get to know more (if she would just answer to my messages...).

To my new family, the Norzi,who welcomed me with the utmost enthusiasm and kindness. To Walter and Elisabetta, who constantly fight against my impostor syndrome and allow me to be surrounded by art. To Francesca and Giulia, for their welcoming attitude and the stories we will share.

To my family, the roots and tree of my existence. To my parents, Paola and Roberto, who gave me so much more than what I needed, taught me the love of knowledge and let me the freedom to be whoever I wanted. To my siblings, Antonio and Alessandra: words cannot express how proud I am of who you are becoming and how much I want to keep on growing with you. To Alfonso, Barbara and Jacopo, who will always be next to me. To my grandparents, Nonna Clelia, Nonno Franco and Nonna Giò, who have the strongest sense of family and always will be my inspiration of strength and devotion.

Finally, to Marco, who took care of me when carrots and hummus were my lunch. Thank you for our reality at the speed of sound (a violin sound! Horrible joke, sorry), that is even better than my dreams. One lifetime does not seem enough for the things we want to do together, but we will make it work.

Contents

Abstract

Bayesian methods constitute a popular approach to perform statistical inference and predict phenomena of interest. Surely, part of the popularity of the Bayesian paradigm can be linked to their intuitive core idea: to take advantage of the user's prior knowledge and integrate it in the statistical procedure. The result of this synergy is the posterior distribution, focal point of Bayesian inference and instrument to quantify the uncertainty of the estimation. This thesis collects the work done as a research student on statistical models, built with the tools of Bayesian nonparametrics, to describe power-law distributed data. The motivation of the proposed models are to be found in two different fields of application: complex networks and privacy assessment.

After the introduction provided by chapter 1, in the first half of the thesis I concentrate on the proposal and analysis of models for networks, the mathematical objects that describe relations among entities by representing them as nodes connected through links. Power-laws usually appear in real-world networks as the distribution function of the degrees, namely the number of links of the nodes. The two pieces of work presented belong to the graphex process framework, a flexible generating process which allows to mimic empirically observed networks characteristics. Chapter 2 fits into this framework and extends the original proposal of [Caron and Fox, 2017] to provide a novel modelling approach to describe sparse networks with spatial structure or other covariates. An approximate inference strategy is provided and tested on simulated data. The paper presented in chapter 3 casts light on the asymptotic properties of networks generated under the graphex process, proving the desirable properties of sparsity, power-law degree distributions, clustering and two central limit theorems.

The second half of the thesis is devoted to the development of statistical methods to quantify the risk of disclosure, which arises whenever datasets

with records of individuals are published: an intruder could match the data with prior information and disclose the identity, and therefore the sensitive features, of a person. Chapter 4 develops an estimator to quantify this risk using the Pitman–Yor process, a popular prior on probability distributions that has a distinguishable power-law tail. A closed form posterior distribution of the estimator is provided in a convenient mixture representation, and experiments on both simulated and real data show the effectiveness of the method. Chapter 5 deals with the estimation of the same risk under no distributional assumptions, using a fully nonparametric method. The estimator is extremely easy to understand, fast to compute and has provable guarantees of optimality.

Chapter 6 concludes the thesis with a final summary and some proposals to extend the current work towards new research directions.

Chapter 1

Introduction

"Everything is linear if plotted log-log with a fat magic marker"
(Mar's law) Akin's laws of spacecraft design

1.1 Power-law distributions

The term "power-law" is employed to indicate a relation between a pair of variables where one varies as a power of the other. If one of the variables represents the probability of observing a phenomenon of at least a certain size, and this is roughly proportional to a power of the size itself, then we are talking about power-law distributions. Such distributions have been observed ubiquitously across natural and human phenomena: the size of craters on Pluto ([Scholkmann, 2016]), the distribution of the frequencies of words in a corpus ([Powers, 1998]), the magnitudes of earthquakes, volcanic eruptions, landslides and wildfires ([Danos, 1998]), the distribution of wealth ([Toda, 2012]), links across web pages ([Albert et al., 1999]) and the size of the population of cities ([Gabaix, 1999]). Because of this widespread observation in different contexts and due to their interesting theoretical properties, communities of researchers in probability and statistics have been using power-law distributions for a long time in fields such as extreme value theory ([Resnick, 1987], [Beirlant et al., 2004]), complex networks models ([Van Der Hofstad, 2009], [Voitalov et al., 2019]) and species sampling problems ([Martín and Goldenfeld, 2006]). In

these communities, and in this thesis, a random variable X on the positive real line is said to be distributed as a power-law if and only if

$$\mathbb{P}(X > x) = \ell(x)x^{-(\alpha-1)} \tag{1.1}$$

with $\alpha > 1$ and $\ell(x)$ a function that, for now, we can assume to admit positive constant limit at infinity. In this definition, the term $x^{-(\alpha-1)}$ dominates $\ell(x)$ in the limit, while on finite values $\ell(x)$ can reshape the power-term quite drastically. Definition 1.1 belongs to a wider family of functions, known as regularly varying, which I will explore more in depth in section 1.3.2.

Power-law distributions are usually associated with some distinctive traits. First, the usual summary statistics, such as mean, variance and higher order moments, do not represent well samples from the distribution or might even be infinite. When the mean value is finite, it might not describe well the observed values, as the occurrences of extreme events (in particular those in the right tail of the distribution) are much more common than in distributions with lighter tails. They have, in fact, heavy tails. Power-laws are also known as "scale-free" distributions, meaning that their behaviour is not affected by different scales, being self-similar along all the domain (for a more detailed account, see [Caldarelli, 2007]). This is shown formally by observing that $\mathbb{P}(X > kx) = \ell(kx)k^{-(\alpha-1)}x^{-(\alpha-1)}$, for k positive real constant: $k^{-(\alpha-1)}$ is a constant and $\ell(kx)$ has the same behaviour of $\ell(x)$ in the limit (since it goes to a constant), hence the dominant term in the limit is again $x^{-(\alpha-1)}$.

In this thesis, I will explore some approaches to describe power-laws in different contexts. Since the first claims about the scale-free behaviour of the degree distributions of real-world networks (the most famous surely regards the Internet graph, [Faloutsos et al., 1999] [Albert et al., 1999]), the claims about the power-law nature of degree distributions have been countless. This is attributed to the fact that some generating processes for networks, which have been identified to mimic empirically observed processes of aggregation, lead to power-law degree distributions. Moreover, scale-free degree distributions are

linked to some macroscopic properties often observed in real data, which will be explored in the next section. This thesis offers a novel approach to describe sparse spatial networks with power-law degree distributions and an extensive analysis of such distributions in a specific class of network generating processes. The second context is statistical disclosure risk limitation, the vast field of statistical studies on disclosure risk assessment, privacy-preserving methods and trade-offs between level of privacy and loss of information. I will focus on the quantification of disclosure risk for data at individual level, which requires to explore the structure of an unobserved sample from a discrete distribution. When it comes to parametric approaches to model such distributions, power-laws are once again a well documented choice ([Martín and Goldenfeld, 2006]).

1.2 An overview of the literature

In this section I will present a general review of the fields of statistical network models and disclosure risk limitation. The section aims at letting the reader dive into the applied contexts that motivated the research presented in this thesis. Since the papers in the following chapters will be more specific, I would like this to be the space to review the broader context and not only the topics strictly connected to my work, which will nevertheless have a special consideration.

1.2.1 Network properties and spatial networks

Graphs are the mathematical construction created to describe relationships among agents. The agents are called nodes or vertices, and when two nodes are in a relationship they are connected by an edge. Nodes could represent any entity, for example humans, computers, chemical elements or cities, and edges could describe relationships such as friendship, wireless connections, chemical bonds or train lines. In more applied contexts, the term "network" is often used in lieu of "graph". In this thesis I will use both terms exchangeably.

One of the most studied network characteristics is the degree distribution,

where the degree of a node is the number of vertices connected to it (referred to as neighbours). Many real-world networks have been observed to display a heavy-tailed degree distribution which was almost often described to be a finite sample from a power-law random variable. Despite the incredible number of examples, though, power-law distributions prove themselves to be quite divisive among researchers in the network modelling community ([Clauset et al., 2009], [Stumpf and Porter, 2012], [Broido and Clauset, 2019], [Voitalov et al., 2019]). Power-laws are measured in their asymptotic behaviour, and this makes finite samples difficult to distinguish from similar data generated from other heavy tailed distributions (for example, the log-Normal). Having a sufficiently large pool of data, covering at least a couple of orders of magnitude, is a necessary requirement to run goodness-of-fit tests. As a result of this, claims about the distributional nature of some benchmark network datasets have been reassessed several times (see, for example, [Willinger et al., 2009] for the internet and [Golosovsky, 2017] for citations). While I will use in this thesis the definition of power-law distribution as eq. (1.1), a more common yet stricter definition, which has often been used to test the networks' degrees, is $\mathbb{P}(X > x) = cx^{-(\alpha-1)}$, where $c > 0$ is the normalising constant and x is greater than a certain value $x_{\min}$. The use of this definition, which does not allow for any flexibility around the pure power function $x^{-(\alpha+1)}$, causes difficulties in the assessment of real network data, noisy and in sample sizes that often are not big enough, and leads to the rejection of the hypothesis of power-law in many goodness-of-fit tests. Instead, due to the presence of $\ell(x)$, definition (1.1) allows for a greater flexibility in the finite regime.

Despite being a fundamental feature, the degree distribution is not the only characteristic of interest. I will introduce other network properties here, but for a more detailed account I refer to [Newman, 2010]. Sparsity is an attribute observed in many real-world networks. The concept of sparsity is related to the relative number of edges with respect to the maximum number of connections achievable, which is proportional to the squared number of nodes, when the network size increases. In particular, when the number of edges grows as a

little-o of the squared number of nodes, the graph is sparse. To picture the concept, think about your favourite social network: despite how popular you might be, the number of people you are not connected with greatly outnumbers the number of friends you have.
Another important feature of real-world networks is clustering, a measure of transitivity in graphs, which quantifies if two nodes with a neighbour in common have higher or lower probability of being connected with respect to the chosen null model. In friendship networks, for example, having a friend in common raises the probability of becoming acquainted. Many social networks display this behaviour, which is measured by a positive clustering coefficient. Also, the vast majority of real network data display the small-world effect. This describes the fact that the the number of edges that need to be traversed to connect two nodes selected at random is surprisingly small, even in huge networks. Formally, to be a small-world network the average number of steps to travel between nodes has to grow proportionally to the logarithm of the number of nodes.

A class of networks studied in this thesis is the one identified by the presence of node covariates or latent features, and more specifically the family spatial networks. An example of such graphs is the network of airport connections, whose nodes represent airports and whose edges describe flight connections, an example of which can be seen in fig. 1.1. We call spatial networks those graphs that lie in a metric space, which allows to measure pairwise distances between nodes. The location of a node can either be a concrete information such as the longitude and latitude of a point on the terrestrial surface, or it could be a latent information. In fact, when there are no geographical coordinates or we are dealing with features of different nature (continuous, discrete, categorical, qualitative...), a distance function might not exist and embedding the network into a metric space would associate a latent coordinate to each node and allow to quantify their similarity through a properly defined distance. For example, when talking about friendship networks, we might not have the spatial information, but we could measure the proximity of individuals

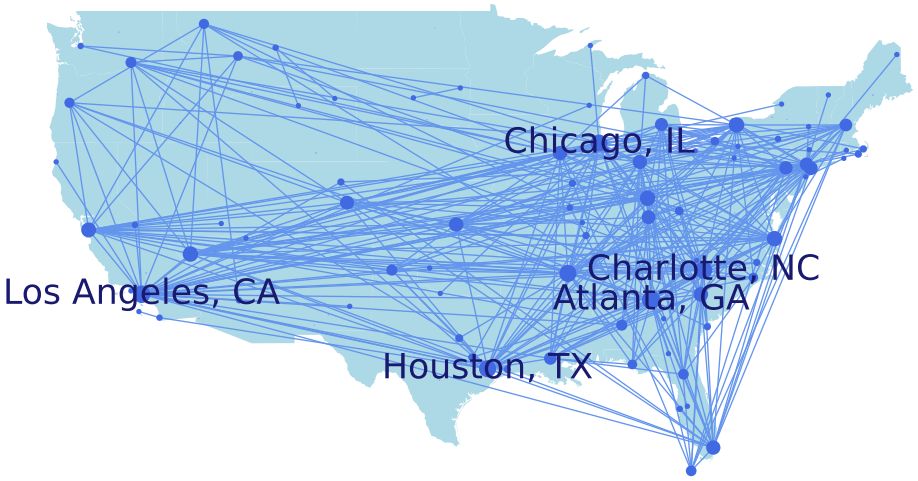


Figure 1.1: A subsample of the US airports connections data in 2010 (excluding Hawaii and Alaska). The dataset comes from `http://toreopsahl.com/datasets/`. The size of the node corresponds to the degree, and the airports with highest degree are highlighted with their names.

in terms of interests, studies or jobs.

1.2.2 Statistical network models

Because of the complex nature of the phenomena described by networks, the collection and storage of network data has seen a huge increase only in the last two decades thanks to advances in the computing technology. This has triggered the widespread interest of quantitative research communities, who have worked to provide models that could describe, understand and predict this type of data. Measuring the network properties described so far does not

let us understand why these behaviours emerge or how they are related. To do so, we want to construct models that allow us to generate multiple observations of graphs having a certain characteristic, study their theoretical properties or design hypothesis tests for them. Random graph models are tools to answer such questions. I will present a brief overview of some of them, and refer for more details to [Newman, 2010].

The first random graph model was proposed more than 60 years ago in [Erdös and Rényi, 1959], who constructed networks by fixing a number of nodes n and sampling edges between each pair of vertices with a probability $p \in [0, 1]$. This structure stimulated the exploration of many theoretical properties, but was too simple to describe real data. For example, its degree distribution does not resemble empirical observations (it is a Poisson distribution in the limit, very far from a scale-free) and its clustering converges to 0 as the number of nodes increases.
To overcome the former challenge, a big advancement in random graph models was the family of configuration models. These models fix a degree sequence and construct a network that can achieve it. Therefore, they can obtain any degree distribution (including the celebrated power-law) and under some conditions they are proven to be small-world ([Chung and Lu, 2002]). Nevertheless, their clustering coefficient is still converging to 0 as the size of the network grows. Other properties of configuration models have been studied in [Molloy and Reed, 1995], [van der Hoorn and Olvera-Cravioto, 2018] and [Kryven, 2017]. Configuration models introduce heterogeneity in the probability of connection between nodes, since the probability of connection is a function of the degree of the nodes. This mechanism, though, does not account for the presence of communities in a graph: blocks of nodes that are highly interconnected among them, but display less connections with the outside.
The stochastic block model (SBM) by [Holland et al., 1983] was developed to describe these communities, defining for each node a probability of connection that varies according to the membership of the other extreme of the edge: higher for nodes in the same community of the starting node, lower for the

rest. Many modifications of the original model have been proposed. I only mention the work by [Karrer and Newman, 2011], who developed a version of the SBM which leads to a better fit of real data, and that of [Airoldi et al., 2008], who proposed a model with overlapping communities.
Another approach to describe networks was proposed by [Watts and Strogatz, 1998], who proposed the small-world model, able to achieve strictly positive clustering and small distances between pairs of nodes.
The cumulative advantage model ([de Solla Price, 1965]), popularly known as preferential attachment ([Barabási and Albert, 1999a]), is the modelling transposition of the "rich get richer" phenomenon. To study the network of citations in academic communities, Price drew insights from [Simon, 1955], who studied the evolution of wealth accumulation: individuals tend to gain an amount of wealth proportional to what they have invested, resulting in an even starker inequality between rich and poor people. The generative process of preferential attachment, which adds at each time step a new node and connects it to a possible neighbour with probability proportional to its degree (hence, the already rich node becomes richer), mirrors this type of growth and induces a power-law degree distribution.
Sometimes the formation and evolution of a network is determined by an optimisation problem determined by the field of application. For example, supply chains and airlines construct their networks balancing the need to save money (i.e. minimising the number of edges) and keep customers happy (minimising the time to traverse the network). Generative mechanisms that try to solve this kind of trade-offs are called network optimisation models, and some examples can be found in [Solé et al., 2003], [Gastner and Newman, 2006], [Aldous, 2008] and [Liu and Zhao, 2012].
The class of exponential random graphs was created to address the issue of statistical inference on networks using the flexible exponential family of distributions. Exponential random graph models prescribe a way to sample from a distribution over a family of graphs which possess a statistic of interest (for example, a predetermined expected number of triangles). There exist some hindrances that limit the applicability of this random graph model: in many

cases, the realisations are either very dense (almost complete) or empty. For an account of the degeneracy problems, see [Chatterjee and Diaconis, 2013], and for general examples and reviews refer to [Holland and Leinhardt, 1981], [Strauss, 1986], [Lusher et al., 2013], [Harris, 2013].

The last model for graphs I am describing is the graphon, whose characteristics and limitations will motivate part of this thesis. In recent times, the size of networks has increased in a vertigious way: the Internet network crossed the threshold of 1 billion web pages in 2014[1], and at the beginning of 2020 Facebook had more than 2.5 billion users. It is not easy to keep track of the sizes of these networks, as they are constantly growing or shrinking and developing or destroying new connections. This is why researchers became more and more interested in models that could handle large and evolving graphs, and study their properties in the asymptotic regime. Motivated by this, limit models for graphs have been explored. The graphon was introduced in [Lovász and Szegedy, 2006] and [Borgs et al., 2017] as the limit of dense graph sequences, namely graphs whose number of edges grows as the squared number of nodes. The graphon made its break-trough in the Bayesian community when [Diaconis and Janson, 2008] drew its connection with the notion of exchangeability for random matrices ([Aldous, 1985a]), which underpins a part of the Bayesian modelling framework for networks. Chapters 2 and 3 are built on an extension of this theory, known as graphex process, and therefore I will review the graphon more thoroughly in section 1.3.3.

1.2.3 Statistical disclosure risk limitation

A risk of disclosure arises whenever an institution wants to publish a file of microdata, i.e. data whose observations are at individual level, for example census or surveys. Having access to data is a fundamental principle for the advancement of analysis and research which will benefit the society, but on the other side of the medal we have the moral imperative of preserving the privacy of those in the dataset. A malicious intruder, for example, could be able to

[1] `https://www.internetlivestats.com/`

match the published data with prior information or other datasets already available and disclose the identities, and therefore the sensitive attributes, of some individuals. This intrusion would not only cause a violation of privacy, but also destroy the confidence of people participating to data collections, and therefore providing reliable estimation of this risk is pivotal. Contrary to the area of network science, which has seen a huge increase of attention only in the last twenty years, the field of data confidentiality has been explored for almost half a century, thanks to the easier collection of data and the huge societal impact. Different types of risk of disclosure exist, leading to diverse methods to tackle them. [Matthews and Harel, 2011] and [Willenborg and Waal, 2001] distinguish among two main categories: risk of re-identification and risk of predictive disclosure, the former focused on measuring the possibility of disclosing the identity of individuals by matching records with their owners and the latter on estimating with accuracy the values of sensitive attributes, even without performing an exact match.

As explained in [Willenborg and Waal, 2001], an intruder could infer the predictive distribution of a sensitive attribute (let us say, income) from the microdata by regressing the variable of interest on the rest of the variables (say gender, age, job and residence). This constitutes an example of predictive disclosure. [Duncan and Lambert, 1989] quantify the knowledge gain that a potential intruder gets from the release of microdata as the relative difference between the uncertainty in the posterior predictive distribution and the uncertainty in the prior predictive distribution. Exceeding a threshold on this relative gain in information would flag the dataset and require further measures to be taken.

Differential privacy is another important player in the context of predictive disclosure. [Dwork, 2006] and [Dwork et al., 2006] shift from an absolute definition of privacy to a relative one, according to which the risk that somebody takes on by sharing their data should be the same as the risk of another individual not participating to the data collection. Take a randomised function $\mathcal{K}$, i.e. a function that takes as argument a dataset and outputs a randomised version

of it. As defined in [Dwork, 2006], a randomised function $\mathcal{K}$ is $\epsilon-$differentially private ($\epsilon > 0$) if for all microdata D_1, D_2 differing in at most one observation and for all sets S in the image of $\mathcal{K}$:

$$\mathbb{P}(\mathcal{K}(D_1) \in S) \leq e^{\epsilon}\mathbb{P}(\mathcal{K}(D_2) \in S).$$

Since the ratio between the two probabilities is bounded from above, the decisions based on the two versions of the dataset (one containing the data of a certain individual and one not) have almost the same probability of happening. If the query requested about some data is the number of 1s in the dataset, an example of randomised $\epsilon-$differentially private mechanism could be to release a value sampled from a Laplace distribution with mean the true number of rows and variance $1/\epsilon$ (the proof can be found in [Dwork et al., 2006]). The field of differential privacy has seen an incredible amount of research attention in the last 15 years. Yet, the number of critiques to the method is not negligible. Models that claimed to be differentially private were discovered not to be so ([Ding et al., 2018]) and there has been a general call for transparency in the applied fields ([Dwork et al., 2019]). Many of the critiques revolve around the difficulty in transposing the results obtained at academical and research level to the practical applications requested by companies and national statistical services. [Garfinkel et al., 2018] explored the practical challenges regarding the choice of the parameter ϵ and of the randomised mechanism, the need of powerful computing resources and well trained personnel and the expectation of the users that will work on the released data. In October 2018, the Integrated Public Use Microdata Series (IPUMS), the largest database of microdata in the world, criticised the decision of the US Census Bureau to implement differential privacy techniques on US microdata, claiming that the statistically-safe usability of public data would have been in danger. More than 4000 academics, researchers and organisations signed their request to delay the start of this procedure. Eventually, the US Census Bureau pushed back the implementation of such mechanisms to 2025, but went on with the decision to pursue the release of the 2020 Census data employing differentially private randomisations, not without criticism ([Kenny et al., 2021] and [Mueller

and Santos-Lozada, 2022] studied the effects of the latest disclosure avoidance system used by the Bureau on the 2010 Census data and reported flawed results in the estimation of the individuals' race and ethnicity).

Setting aside predicting disclosure and focusing on the methods to tackle the risk of identification, k-anonimity is commonly used by companies dealing with sensitive data[2]. This method ([Samarati and Sweeney, 1998], [Sweeney, 2002]), prescribes that each record in the microdata should not be distinguishable from at least other $k-1$ individuals in the sample. If the original data do not satisfy it, $k-$anonimity is obtained by suppression or generalisation. Suppression, as the name suggests, prescribes to cancel the entries of some cells, while the generalisation technique groups categories of an attribute into a coarser classification (for example, instead of the city it could show the province). Despite the apparent simplicity, choosing how to modify the dataset is an incredibly difficult problem ([Meyerson and Williams, 2004]) and new algorithmic proposals have been explored to circumvent it (for example, [Bayardo and Agrawal, 2005], [Kenig and Tassa, 2012]).
Another stream of disclosure methods originated with [Bethlehem et al., 1990], who identified as participants at highest risk those that appeared only once in the microdata sample and were also unique in the underlying population. From now on, I will identify the number of sample uniques which are also population uniques as τ_1. In table 1.1 I provide an illustrative toy example to understand this quantity. Estimating τ_1 can help in quantifying the risk associated with a specific dataset and address the question of publication, for example by imposing a relative threshold over which the data will not be released or will be modified using other privacy-preserving methods. The challenge of this approach is the need to estimate the unobserved part of the population, and different ways to do it were explored in [Samuels, 1998], [Skinner and Elliot, 2002a], [Rinott and Shlomo, 2006], [Carota et al., 2015]. The estimation of τ_1 is the framework on which I have worked for part of this thesis, in chapters 4 and 5.

[2] `https://centre.humdata.org/learning-path/disclosure-risk-assessment-overview/`

		Gender	# Kids	Education	Residence	
Sample	□	F	1	Degree	Oxford	Population
	⋆	M	7	PhD	Birmingham	
	□	F	1	Degree	Oxford	
	□	F	1	Degree	Oxford	
	●	F	3	Diploma	Manchester	
	●					
	⊗					

Table 1.1: Example of microdata file, where the first 5 rows correspond to the observable sample and the last 2 to the rest of the population. The symbols on the left uniquely identify the possible records (there are 4 in the population, and 3 appear in the sample). In this case, the number of sample uniques which are also population uniques τ_1 is 1, corresponding to the individual identified by ⋆.

The estimation of τ_1 is directly linked to species sampling problems, a broad statistical question that looks at the estimation of functionals of discrete distributions regarding unobserved samples. A long-standing example is the estimation of the number of unseen species ([Fisher et al., 1943], [Goodman, 1949], [Good, 1953], [Good and Toulmin, 1956], [Efron and Thisted, 1976], [Orlitsky et al., 2016]). The most prolific field of application of species sampling problems is biology, where they have been used for example to estimate variants in the genome ([Ionita-Laza et al., 2009]), complexity of genomic sequencing ([Daley and Smith, 2013]) and bacterial presence ([Gao et al., 2007]). Other applications have been explored, such as unseen words in a corpus ([Efron and Thisted, 1976]) or password habits [Florencio and Herley, 2007]. In these examples, the population is sampled from a discrete distribution: there are a countable (possibly infinite) number of species, and each of them has a probability p_j of being observed, such that $\sum_{j\geq 1} p_j = 1$. The estimation procedures of such distributions can be fully nonparametric, avoiding any assumption on the mechanism originating the data, or impose some parametric shape on it. When it comes to parametric approaches, power-law distributions are once again a documented choice for this problems ([Martín and Goldenfeld, 2006], [Gnedin et al., 2007]).

1.3 Background

The vast majority of the methods in this thesis are built under the Bayesian framework. A Bayesian model is defined through the specification of an a priori belief over the parameter θ, the prior distribution $\pi(\theta)$, and a function describing the likelihood of observing some data points x conditioned on θ, the likelihood function $p(x|\theta)$. Employing Bayes theorem, the updated uncertainty over the parameter is obtained in the form of a posterior distribution over θ conditioned on the observations x:

$$\pi(\theta|x) \propto p(x|\theta)\pi(\theta).$$

When the dimension of the parameter θ is finite, parametric Bayesian inference is used. In this thesis, I will use Bayesian nonparametric methods, designed for infinite dimensional parameters. For a comprehensive review of Bayesian inference, see [Gelman et al., 2003] and [Robert et al., 2007]. Specifically on Bayesian nonparametrics, see [Ghosal and Van der Vaart, 2017] and [Hjort et al., 2010].

1.3.1 Completely random measures

Completely random measures (CRMs) are a building block of Bayesian nonparametric theory, being one of the tools to construct prior distributions over functional spaces and discrete random structures. Firstly, I will revisit the concept of Poisson Process, a famous atomic random measure and crucial ingredient of CRMs.

Definition 1 (Poisson process. [Kingman, 1993])**.** *A Poisson process N on a measurable space $(S, \mathcal{S})$ is a random measure such that for every collection of measurable disjoint sets $A_1, \ldots, A_n \in \mathcal{S}$ it has $N(A_i) \overset{ind}{\sim} Poisson(\nu(A_i)), i = 1, \ldots, n$, where ν is a measure over $(S, \mathcal{S})$. ν is called the mean (or intensity) measure of the process. We identify such Poisson process as $N \sim Poisson(\nu)$.*

Definition 2 (Completely random measure [Kingman, 1967])**.** *A completely random measure μ over a measurable space $(S, \mathcal{S})$ is a random measure on*

the same space such that for every disjoint measurable sets A and B, $\mu(A)$ is independent of $\mu(B)$.

A characterization theorem for CRMs is available, which highlights the connection with Poisson processes.

Theorem 3 (Theorem 4 [Kingman, 1967])**.** *A CRM μ on $(S, \mathcal{S})$ can be decomposed as the sum of three independent components:*

$$\mu = \mu_0 + \sum_{j=1}^{\infty} v_j \delta_{x_j} + \int w N(dw, d\theta) \tag{1.2}$$

where μ_0 is a purely deterministic measure, $\sum_{j=1}^{\infty} v_j \delta_{x_j}$ is an atomic random measure with $(v_j)_j$ independent random variables on $\mathbb{R}_+$ and $(x_j)_j$ a countable set of fixed atomic locations in S. The last term of eq. (1.2) *is derived from a Poisson process N on $\mathbb{R}_+ \times S$ with mean measure $\nu(dw, d\theta)$ which satisfies the following condition:*

$$\int_B \int_{\mathbb{R}_+} \min\{w, 1\} \nu(dw, d\theta) < \infty$$

on any measurable set $B \subset \mathcal{S}$. ν is referred to as the Lévy measure of the CRM.

Of special interest in this thesis are homogeneous CRMs, a specific class of CRMs whose Lévy measure can be factorised into two components.

Definition 4 (Homogeneous CRM)**.** *Consider a CRM μ as in eq.* (1.2)*. When $\nu(dw, d\theta)$ can be decomposed into independent components $\rho(dw)H(d\theta)$, then μ is called homogeneous.*

Following [Caron and Fox, 2017], for the rest of this work I will concentrate on homogeneous CRMs on $\mathbb{R}_+ \times S$ uniquely represented by the third term of eq. (1.2). Such CRMs can be equivalently written as

$$\sum_{i=1}^{\infty} w_i \delta_{\theta_i} \tag{1.3}$$

with Lévy measure $\rho(dw)H(d\theta)$.

Examples of completely random measures

The generalised gamma process (GG) ([Hougaard, 1986], [Aalen, 1992], [Lee and Whitmore, 1993], [Brix, 1999]) is an example of family of CRMs, well studied due to its computational tractability. It is a homogeneous CRM of the type of eq. (1.3) with Lévy measure

$$\rho_{GG}(dw)H(d\theta) = \frac{1}{\Gamma(1-\sigma)} w^{-\sigma-1} e^{-cw} dw H(d\theta) \tag{1.4}$$

with $(\sigma, c) \in (-\infty, 0] \times (0, \infty)$ or $(\sigma, c) \in (0, 1) \times [0, \infty)$. Noteworthy subcases are the gamma process obtained with $\sigma = 0$ and $c > 0$, the stable process with $\sigma \in (0, 1)$ and $c = 0$, and the inverse-Gaussian process with $\sigma = 1/2$ and $c > 0$ ([Lijoi et al., 2005]).

Another example is the generalised gamma Pareto process (GGP), introduced in [Ayed et al., 2019] as an extension of the BFRY distribution by [Bertoin, 2006] and [Devroye, 2009]. It is characterised by the Lévy measure

$$\rho_{GGP}(dw)H(d\theta) = \frac{1}{\Gamma(1-\sigma)} w^{-\sigma\tau-1} \gamma(\sigma(\tau-1), cw) dw H(d\theta) \tag{1.5}$$

where $\sigma \in (-\infty, 1), \tau > 1, c > 0$ and $\gamma(k, y) = \int_0^y u^{k-1} e^{-u} du$ is the lower incomplete gamma function. Samples from the GG and GGP have a distinct power-law behaviour, and I will illustrate it in section 1.3.2.

CRMs have been widely used in Bayesian nonparametrics as a tool to construct priors over a set of distributions (see, for example, [Ghosal and Van der Vaart, 2017], [Hjort et al., 2010]). In this case, one needs to normalise a CRM (whose total mass is tipically not 1) to obtain a normalised completely random measure (NCRM). A notorious example of NCRM is the Dirichlet process ([Ferguson, 1973]). Normalisation is not the only way to obtain such priors, and some alternative constructions have been proposed that also enhance the understanding and interpretation of CRMs. One of such examples is given by the construction of the Pitman–Yor process (of which the Dirichlet is a

subcase): the stick-breaking representation proposed in [Pitman and Yor, 1997]. For $\alpha \in [0, 1)$ and $\theta > -\alpha$, let $(z_j)_{j\geq 1}$ be random variables independently and identically distributed from a non-atomic distribution on a space S. Let $(v_i)_{i\geq 1}$ be independent of the $(z_j)_j$ and distributed according to the following rule:

$$v_i \overset{ind}{\sim} \text{Beta}(1-\alpha, \theta + i\alpha).$$

Setting

$$\begin{cases} p_1 = v_1 \\ p_j = v_j \prod_{i=1}^{j-1}(1-v_i) \quad j > 1 \end{cases} \tag{1.6}$$

ensures that $\sum_{j\geq 1} p_j = 1$ almost surely. Then

$$P_{\alpha,\theta} = \sum_{j\geq 1} p_j \delta_{z_j}$$

is a Pitman–Yor process on S with discount parameter α and scale parameter θ. The Dirichlet process arises as a sub-case by letting $\alpha = 0$.
The Pitman–Yor process is as well linked to the power-law phenomenon. In fact, for $\alpha \in (0, 1)$, let $(p_{(j)})_{j\geq 1}$ be the random probabilities p_j's of eq. (1.6) in decreasing order. Then, as j grows to infinity (and $p_{(j)}$ decreases), $p_{(j)}$ has power-law behaviour with exponent α^{-1}. The Pitman–Yor process generalizes the Dirichlet process by means of the "discount" parameter α which controls the tail behaviour of $P_{\alpha,\theta}$, ranging from a geometric tail (in the case $\alpha = 0$ for the Dirichlet process) to a heavy, power-law tail (the higher α, the heavier the tail).

1.3.2 Regular variation

As motivated in section 1.1, the definition of power-law distribution used in this thesis will correspond to eq. (1.1). Such definition belongs to the world of regularly varing functions, of which I am about to give the formal definition. Examples and connections with CRMs will follow.

Regularly varying functions are characterised by a limiting behaviour (at infinity or 0) that resembles a power law with an additional slack. For a detailed account, see [Bingham et al., 1987].

Definition 5 (Slow variation, [Bingham et al., 1987])**.** *A positive function f on $(0, \infty)$ is said to be slowly varying at infinity if $\lim_{y\to\infty} f(ty)/f(y) = 1$ for any $t > 0$.*

Examples of slowly varying functions are $\log^a x, a > 0$, or any function with positive constant limit at infinity.

Definition 6 (Regular variation, [Bingham et al., 1987])**.** *A positive function f is regularly varying at infinity with exponent α if it can be written as $f(y) = y^\alpha \ell(y)$, with $\ell(y)$ slowly varying at infinity. f is regularly varying at 0 with exponent α if $f(1/y)$ is regularly varying at infinity with exponent $\alpha \in \mathbb{R}$, or equivalently $f(y) = y^{-\alpha}\ell(1/y)$.*

Intuitively, with this definition we are allowing a power function to be smudged by a slowly varying function, which does not change too quickly at infinity. Note that a function can be regularly varying at both extremes with different exponents.

In this work, the interest is focused on regularly varying CRMs. This translates in the analysis of regular variation measured on their respective tail Lévy measures $\bar{\rho}(y) := \int_y^\infty \rho(dw)$. For two functions f and g, let $f \sim g$ indicate that $\lim_{x\to\infty} f(x)/g(x) = 1$. In the case of the GG of eq. (1.4), as y approaches 0 $\bar{\rho}$ is

$$\bar{\rho}_{GG}(y) \sim \begin{cases} \frac{1}{\Gamma(1-\sigma)\sigma} y^{-\sigma} & \text{for } \sigma > 0 \\ \log(1/y) & \text{for } \sigma = 0 \\ -c^\sigma/\sigma & \text{for } \sigma < 0 \end{cases} \tag{1.7}$$

Hence, the GG is regulary varying at zero with exponent $\max\{0, \sigma\}$. Due to the exponential decay at infinity of eq. (1.4), the GG is not regularly varying at infinity.

The GGP of eq. (1.5) has tail intensity such that

$$\bar{\rho}_{GGP}(y) \sim \begin{cases} c_1 y^{-\sigma} & \text{as } y \to 0 \\ c_2 y^{-\sigma\tau} & \text{as } y \to \infty \end{cases} \tag{1.8}$$

for some positive real constants c_1, c_2. Therefore, the GGP is regularly varying both at infinity and at 0, with exponents $-\sigma\tau$ and σ respectively.

1.3.3 Graphon and graphex

An important concept in Bayesian statistics is that of exchangeability, which intuitively describes a relationship among random variables such that the order of appearance does not matter. In many settings, this invariance with respect to an order is a very reasonable assumption, and in statistical models it is desirable because of its implications of computational and theoretical tractability. I will present here more formally the concept of exchangeability and some representation theorems, and for more details refer to [Aldous, 1985b].

Definition 7 (Exchangeable sequence [Aldous, 1985b])**.** *Let $(X_i)_{i\geq 1}$ be an infinite sequence of random variables taking values in a space $\mathbf{X}$ with $\sigma-$ algebra $\mathcal{X}$. The sequence is exchangeable if*

$$\mathbb{P}(X_1 \in A_1, X_2 \in A_2, \dots) = \mathbb{P}(X_{\pi(1)} \in A_1, X_{\pi(2)} \in A_2, \dots) \tag{1.9}$$

for every collection of sets $A_i \in \mathcal{X}$ and any permutation π over $\mathbb{N}_+ := \{1, 2, \dots\}$.

The structure of exchangeability is linked to a representation theorem. In the case of the sequences just described, this was proposed in [de Finetti, 1931], [De Finetti, 1937].

Theorem 8 ([De Finetti, 1937])**.** *A sequence $(X_i)_{i\geq 1}$ is exchangeable if and only if*

$$\mathbb{P}(X_1 \in A_1, X_2 \in A_2, \dots) = \int_{P_{\mathbf{X}}} \prod_{i\geq 1} p(A_i)\nu(dp)$$

with $P_{\mathbf{X}}$ the space of probability measures on $(\mathbf{X}, \mathcal{X})$, $\mathbf{X}$ complete and measurable and $\mathcal{X}$ its Borel $\sigma-$algebra, and ν a probability measure over $P_{\mathbf{X}}$.

In words, this theorem means that exchangeable sequences are mixtures of independent and identically distributed random variables. Therefore, in Bayesian inference exchangeable sequences are obtained by first sampling a probability measure p from the prior distribution ν, then sampling the observations from the likelihood p.
To describe networks, though, the concept of sequence has to be extended. The most employed and easiest representation of a graph is through its adjacency matrix Z, a square binary matrix representing the pairwise connections between nodes: if an edge is present between nodes i and j, the entry Z_{ij} will be 1, otherwise 0. From now on, I will use Z to indicate binary variables and X to indicate any variable (not necessarily binary).

Definition 9 (Exchangeable matrix [Aldous, 1985b])**.** *Let $X = (X_{ij})_{i,j\geq 1}$ be a matrix of random variables and define $R_i := (X_{ij}, j \geq 1)$ the sequence of row vectors and $C_j := (X_{ij}, i \geq 1)$ that of columns. X is jointly exchangeable if both R_i and C_j are exchangeable sequences according to definition 7, for any permutation π applied to both R and C.*

In the context of random graphs, this type of exchangeability is known as node exchangeability, as it corresponds to a reshuffling of the nodes according to a given permutation π. The representation theorem for random matrices was proposed in [Aldous, 1981] and [Hoover, 1979].

Theorem 10 ([Aldous, 1985b])**.** *A matrix $X = (X_{ij})_{i,j\geq 1}$ is jointly exchangeable if and only if there exists a random function $f : [0,1]^3 \to \mathbf{X}$ such that*

$$(X_{ij})_{ij} \stackrel{d}{=} f(U_i, U_j, U_{ij}) \tag{1.10}$$

where $(U_i)_i$ and $(U_{ij})_{ij}$, $U_{ji} = U_{ij}$ are respectively a sequence and a matrix of independent and identically distributed Uniform$[0,1]$ random variables, independent of f.

[Diaconis and Janson, 2008] opened the door for a connection between the Aldous-Hoover theorem 10 and the graphon function, the limit of dense graph

sequences introduced in section 1.2.2 ([Lovász and Szegedy, 2006]; [Borgs et al., 2010]). Consider a jointly exchangeable graph without loops (edges from a node to itself); according to theorem 10 there exists a function f satisfying eq. (1.10) which maps into $\mathbf{X} = \{0, 1\}$. Then, the graphon $\mathcal{W}$ is defined as the symmetric, measurable function from $[0, 1]^2$ to $[0, 1]$ such that

$$\mathcal{W}(x, y) := \mathbb{P}(f(x, y, U) = 1|f) = \int_0^1 f(x, y, u)du$$

with U distributed as a Uniform$[0, 1]$ and $\mathcal{W}(x, x) = 0$. From this perspective, it is possible to rewrite theorem 10 specifically for exchangeable random graphs (see Corollary III.6 in [Orbanz and Roy, 2015]). Any node-exchangeable network with adjacency matrix Z is represented by a graphon function $\mathcal{W}$ such that

$$Z_{ij}|(\vartheta_k)_{k=1,2,\dots} \sim \text{Bernoulli}(\mathcal{W}(\vartheta_i, \vartheta_j)) \tag{1.11}$$

where $(\vartheta_k)_k$ are independent and identically distributed Uniform$[0, 1]$. And the reverse holds true as well, since every symmetric measurable function from $[0, 1]^2$ to $[0, 1]$ describes a node-exchangeable network.
This representation is very powerful, but hides an important cost: a consequence of theorem 10 is that networks generated under this framework are either empty or dense. In fact, since in a graph with n nodes there are up to $\binom{n}{2}$ edges, the expected proportion of edges in such network is given by $\binom{n}{2}\frac{1}{2}\epsilon$, with $\epsilon := \int_{[0,1]^2} \mathcal{W}(x, y)dxdy$. As n grows to infinity, this quantity can be either 0 if ϵ is 0, or $\Theta(n^2)$ if ϵ is positive, making the graph respectively dense or empty and the model inherently misspecified for real graphs which are usually sparse.

Some approaches to tackle this problem have been explored. Models like preferential attachment ([Barabási and Albert, 1999b], [Berger et al., 2014]) or configuration ([Bollobás, 1980], [Newman, 2010]) achieve sparsity by setting aside exchangeability. Another possibility is to rescale the graphon function so that the probability of connection becomes smaller as the size of the graph grows ([Bollobás and Riordan, 2009], [Borgs et al., 2019]). These models

are finitely exchangeable but lack in projectivity, the property assuring that inference performed on a graph of dimension n remains coherent when moving to a version of the same graph with more nodes. Another possibility is to rely on edge exchangeability instead of node exchangeability, as done in [Crane and Dempsey, 2015], [Broderick and Cai, 2016], [Williamson, 2016], [Janson, 2017].

A different solution was advanced by [Caron and Fox, 2017] who, instead of using the adjacency matrix, represented the graph as a point process on the real positive plane:

$$Z = \sum_{i,j \in \mathbb{N}_+} Z_{ij} \delta_{(\theta_i, \theta_j)}, \tag{1.12}$$

where Z_{ij} has the same binary values of the adjacency matrix and $\theta_i \in \mathbb{R}_+$ represents the label of node i. The notion of exchangeability is translated to the world of point processes.

Definition 11 (Exchangeable point process [Caron and Fox, 2017])**.** *A point process Z as eq.* (1.12) *is jointly exchangeable if and only if*

$$X(A_i \times A_j) \stackrel{d}{=} X(A_{\pi(i)} \times A_{\pi(j)}),$$

where $A_i = [h(i-1), hi]$, for any $h > 0$ and any π permutation on $\mathbb{N}_+$.

The representation theorem associated with this definition is the continuous time (and more involved) version of theorem 10 of [Kallenberg, 1990]. As illustrated in [Caron and Fox, 2017], this notion is the keystone that allows their construction to describe the full spectrum of empty, dense and sparse graphs.
[Veitch and Roy, 2015] and [Borgs et al., 2019] expanded the theory around [Caron and Fox, 2017] and showed that this new framework is a generalisation of the graphon framework. They were the first using the term "graphex" for this larger class of random graphs. The properties of graphexes were further studied in [Janson, 2016], [Janson, 2017] and [Borgs et al., 2018]. Following the notation of [Veitch and Roy, 2015], the graphex process is identified by the sparse graphon function, a symmetric measurable function $W : \mathbb{R}_+^2 \to [0, 1]$

such that

$$Z_{ij}|(\vartheta_k)_{k=1,2,\dots} \sim \text{Bernoulli}(W(\vartheta_i, \vartheta_j))$$

where $(\vartheta_k)_{k=1,2,\dots}$ are points of a unit-rate Poisson process on $\mathbb{R}_+$. The sparse graphon W generalises the graphon $\mathcal{W}$ of eq. (1.11) by extending its definition on the whole positive real plane. The graphex process is the modelling framework under which chapter 2 and chapter 3 will be developed.

1.4 Contributions and thesis outline

This thesis comprises four main chapters, each representing independent work. Chapters 2 and chapter 3 focus on networks, the former containing contributions in terms of new Bayesian nonparametric modelling techniques for spatial networks, and the latter proving some asymptotic characteristics of the graphex process. Chapter 4 and 5 detail new statistical methods to quantify disclosure risk, the first in a power-law setting, while the second without distributional assumptions. Chapter 2 has not been submitted, chapter 3 has been submitted to a journal and chapters 4 and 5 have been published. Finally, in chapter 6 I offer a final summary of the whole work and draw some directions for future exploration.

In the next subsections I am going to briefly present the methodologies, results and impact of the following chapters.

1.4.1 Sparse spatial random graphs

In chapter 2, joint work with F. Caron and J. Rousseau, I present an extension of the original base model for sparse random graphs of [Caron and Fox, 2017] in the context of spatial networks. We attach to each node a location living in a metric space and define a connection probability function that allows for a dependence on the distance between nodes that can be tuned with a parameter.
The methodology behind it lies in the world of Bayesian nonparametrics, in

particular in the theory of completely random measures. We obtain interpretable parameters and appealing network characteristics such as a power-law degree distribution, positive clustering coefficients and sparsity. All of these features can be easily tuned. To sample from the model we propose an efficient algorithm that relies on spatial grids to reduce the computational complexity. For the posterior inference of the variables and the parameters, we use the approximation for CRMs proposed by [Lee, 2019] and construct a Markov Chain Monte Carlo algorithm targeting the approximated posterior. We show that this inference scheme works on simulated data. Finally, we discuss our proposal in the context of the broader literature on spatial networks.

We propose this model because we think that the presence of the edges in the base proposal of [Caron and Fox, 2017] can be enhanced by considering nodes' proximity in space. This is true, for example, in transportation networks or social networks, where closeness in space usually corresponds to a higher likelihood of connection. Our model, though, does not need to be specifically considered as spatial in the proper sense of the word. In fact, locations could be the result of the embedding of any network (without concrete spatial information) in a latent metric space, for example when we want to quantitatively measure the affinity among qualitative features of nodes. At the same time, any quantitative covariate (not necessarily a location) could be employed, making our model a flexible option for various applications.

1.4.2 On sparsity, power-law and clustering properties of graphex processes

In chapter 3, I present the joint work with F. Caron and J. Rousseau on some asymptotic properties of the graphex process. In the general graphex framework, introduced in section 1.3.3, we show under which assumptions the resulting graph is sparse or dense and has power-law degree distribution, possibly with different exponents for high and low degrees. We provide the asymptotic values of the global and local clustering coefficients, and derive central limit theorems for the number of nodes and subgraphs. Interestingly,

many of these results rely solely on a regularly varying behaviour of the marginal graphon function μ, defined as

$$\mu(\vartheta) := \int_0^\infty W(\vartheta, y)dy.$$

Intuitively, $\mu(\vartheta)$ is proportional to the expected degree of a node with parameter ϑ and its regular variation translates onto the degree distribution.
After having proved the main results, we illustrate many examples of possible sparse graphon functions, exploring sparse and dense networks and studying more in depth the model by [Caron and Fox, 2017] under different choices for the distribution of the node weights. We also show how the graphex framework can be extended to include local structures such as communities or new covariates (for example space, as proposed in the previous section 1.4.1), and prove some asymptotic results in this more general setting.

1.4.3 Bayesian nonparametric disclosure risk assessment

In chapter 4, I present a way to estimate the number of sample uniques which are also population uniques τ_1, introduced in section 1.2.3. This work is joint with S. Favaro and T. Rigon. The proposed estimator, first introduced in [Cerquetti, 2013], is based on the Pitman–Yor process (presented in section 1.3.1). Under this prior, we are able to compute in closed form the posterior distribution and the expected value of the estimator. In particular, we show how the posterior can be written as a mixture of the hypergeometric and generalised factorial distributions. To provide uncertainty quantification of the estimate, we rely on Monte Carlo techniques to sample from the posterior, whose parameters are estimated using the empirical Bayes approach.

Our work provides an estimator for the number of sample uniques that are also population uniques which is easy to implement and compute, even on massive datasets. Moreover, it has a nice interpretation as the proportion of the uniques in the sample weighted by a quantity easily explained by the Pitman–Yor parameters' values. We prove that our estimator works well when

the underlying distribution over the population has power-law or geometric tails (when the Pitman–Yor collapses to the Dirichlet process), by running experiments that compare different estimators on simulated and real data. By doing so, we show how our estimator is useful to overcome the underestimation of τ_1 which is usually obtained by other approaches in the literature. Alongside, we provide an easy way to empirically validate if the assumption of power-law or geometric tails is satisfied, and therefore to check when our model is correctly specified for the data in question. As real data, we opt for a dataset from the 2018 American Community Survey, a random sample of the American population.

1.4.4 Optimal disclosure risk assessment

In the context of disclosure risk, in chapter 5 I present the work done jointly with F. Camerlenghi, S. Favaro and Z. Naulet on a new estimator for the number of sample uniques which are also population uniques τ_1. We do so by exploiting fully nonparametric techniques which do not impose any distributional assumption on the population distribution (differently from the work just introduced in section 1.4.3) and result in an estimator which is very easy to compute, without the need of any Monte Carlo technique. Our proposal has also an interesting connection with the empirical Bayes approach in the sense of [Robbins, 1956], which can be used to equivalently derive the same estimator.

Chronologically, this paper came before chapter 4, as we first felt the need to solve the problem of estimation of τ_1 under a very general setting. After having finished it, we decided that imposing some distributional assumptions would have been an interesting addition to the literature, and looked at the power-law behaviour as a well motivated choice for it. Moreover, the parametric approach allowed us to perform uncertainty quantification, which is something we are not able to provide in the fully nonparametric setting.

This paper represents an important step in the literature, since it provides the answer to the long standing question of nonparametric estimation of τ_1

posed by [Skinner and Elliot, 2002b]. We close the open problem by proposing a nonparametric estimator which has desirable properties. Firstly, under the assumption of a reasonably sized sample, we show that our estimation procedure is optimal in the sense of vanishing normalised mean squared error with matching upper and lower bounds. Secondly, we prove that when the sample size is too small it is not possible to perform efficient nonparametric estimation (meaning, again, with vanishing normalised mean squared error). From a theoretical point of view, it is worth mentioning that the construction used to bound from below the mean squared error of the estimator exploits complex approximation techniques, since the problem can be rewritten as an estimation of the best polynomial approximation to a certain functional. This represents an advancement in the field of approximation theory, developing for the first time a proof for a more complicated case than what is usually encountered in the literature.

Chapter 2

Sparse spatial random graphs

Unpublished and unsubmitted work.

Sparse spatial random graphs

Francesca Panero

Department of Statistics, University of Oxford
email francesca.panero@stats.ox.ac.uk

François Caron

Department of Statistics, University of Oxford
email caron@stats.ox.ac.uk

Judith Rousseau

Department of Statistics, University of Oxford
email judith.rousseau@stats.ox.ac.uk

Abstract

We present a statistical model for graphs with sociability and spatial components, founded on the theoretical framework of the graphex process and completely random measures. We prove a number of asymptotic results for the number of nodes, edges, degree distribution and clustering properties, showing that the model allows for the description of both dense and sparse networks with a tunable sparsity level and power-law degree distributions for low degree nodes. We offer a time-efficient way to simulate from the model and an approximate Markov chain Monte Carlo algorithm to perform posterior inference. Finally, we show the performance on simulated data and we compare our theoretical framework with other similar spatial network models.

Keywords: Networks, spatial structure, Bayesian nonparametrics, completely random measures, point processes, sparsity, power-law, graphon.

1 Introduction

A graph is a mathematical construction to describe a set of entities, called nodes or vertices, and the pairwise relations among them, identified by edges drawn between nodes. When this modeling construction is translated into a more applied context, the term network becomes more common. In this work we will use both terms with the same meaning. Networks, as many mathematical constructions, possess the ability to describe a huge range of applications, from social networks to biological interactions, from transportation networks to internet connections, from molecule structures to article coauthorships and many more. Due to their complex relational nature, real world network data were hard to collect and store until a few decades ago, but recently the availability of them surged dramatically thanks to the advancement in storage

and computing technology. And as data availability started to increase, so did the interest of the statistical research community.

Starting from Erdös and Rényi (1959) more than 60 years ago, various models have been proposed in probability and statistics to fit these complex relations in the most precise way yet taking into account noise. Empirical analyses in the last decades have highlighted a set of characteristics that seem to be common among real-world data across domains: scale-free degree distribution, that describes the otherwise unexplained presence of nodes with very high degrees and multitude of nodes with small degrees; the small-world phenomenon, according to which every pair of randomly picked nodes is connected through a surprisingly short path; non-vanishing clustering coefficient, the mathematical transposition of the fact that individuals with a friend in common are more likely to become friends; sparsity in the number of connections and presence of densely connected communities. A statistical model has to be flexible enough to encapsulate at least some of these behaviours, but the desiderata of a modelling construction do not end up there. At the same time, we want to achieve both theoretical and computational tractability, that allow us to analyse theoretical properties and scale the inference to big datasets in a limited time. Interpretability of parameters is another desirable property to understand the obtained results. For a broad literature review of network models, we refer to Bollobás (2001), Dorogovtsev et al. (2003), Newman (2003), Caldarelli (2007), Durrett (2007), Cohen and Havlin (2010), Newman (2010), and Fienberg (2012).

In the vast majority of real-world networks, connections are not homogeneous throughout all nodes, but are usually determined by observed or unobserved variables which differ across nodes. For example, in social networks we know that jobs, interests, location, age, gender and other individual variables can influence the connections among agents. A type of heterogeneous network model that has been extensively studied is the spatial graph. Spatial networks, i.e. graphs whose nodes live in a metric space, have been the subject of increasing attention since their emergence in the 1970s. First approached in quantitative geography (Haggett and Chorley (1969)), they have since then evolved to represent a useful model to study various applications such as transportation, mobile phones, power grids, brain connections, the Internet, social studies, epidemiology and others. In all these examples, the network structure cannot be described without taking into account the topology of the space in which the nodes are embedded. For an extensive review about spatial networks see Penrose (2003) and Barthélemy (2011). But while in all these examples space is an observable covariate, this does not always have to be the case. In fact, some networks without a concrete spatial structure might still benefit from being embedded into an auxiliary metric space. The connection between two nodes could be influenced by a similarity among covariates of different nature (discrete, continuous, categorical, with different scales, qualitative...) or even by some unobserved factors, and embedding the nodes into a metric space allows to define a meaningful distance between them, giving rise again to a spatial graph. Latent space models for networks originated in Hoff et al. (2002); for an interesting and convenient embedding of the internet graph see Boguñá et al. (2010).

Various statistical and probabilistic models for spatial networks have been proposed. Among others, the hyperbolic random graph (HRG) by Krioukov et al. (2010) and its Euclidean equivalent, the geometric heterogeneous random graph (GIRG) by Bringmann et al. (2019) and Komjáthy and Lodewijks (2019), proved themselves able to achieve power-law degree distributions, positive clustering and small and ultra small-world behaviours. The scale-free percolation model (SFP) on $\mathbb{Z}^d$ by Deijfen et al. (2013) and its continuum version on $\mathbb{R}^d$ by Deprez and Wüthrich (2018) were introduced to study degree distributions, distances properties and percolation. SFP and GIRG are the spatial counterparts of the proposals previously developed

by Norros and Reittu (2006) and Chung and Lu (2002), where each node i is characterised by a variable w_i, interpreted as sociability, that explains the different degrees associated to each node. These heterogeneous models are able to achieve scale-free degree distributions and small-worldness but cannot obtain positive clustering. Their spatial versions are developed to solve this challenge, since taking into account the topology of the space is a workaround to achieve positive clustering. The sparse latent position model by Spencer and Shalizi (2017) originates from a different stream of models, the latent position models by Hoff et al. (2002). Being purely spatial and with no sociability variables, this model does not describe scale-free degree distributions.

Setting aside space and zooming out on the general frameworks to describe random graphs, the graphon setting (Lovász and Szegedy (2006), Borgs et al. (2010)) is particularly important because it describes the whole class of node exchangeable graphs, i.e. structures where the labels of the nodes do not influence the distribution of the observed connections. Nevertheless, the definition of infinite exchangeability associated with the graphon (Aldous (1981), Aldous (2009)) has been proved to generate only dense or empty graphs (see, for example, Orbanz and Roy (2015)). Networks with sparse connections, though, represent the vast majority of real data and make the original graphon framework misspecified for real world applications. A number of extensions of the graphon have been proposed to overcome this limitation. By setting aside exchangeability, the preferential attachment (Barabási and Albert (1999), Berger et al. (2014)) and the configuration models (Newman (2010), Bollobás (1980)) have been proposed. Bollobás and Riordan (2009) and Borgs et al. (2019) extended the original setting obtaining sparse graphons, able to describe sparse networks but lacking in projectivity, a desirable property to avoid consistency failures in certain estimation problems (Orbanz (2010), Spencer and Shalizi (2017)). Another approach, initiated by Caron and Fox (2017), Borgs et al. (2016) and Veitch and Roy (2015), introduced the graphex process, a framework in which a graph is represented as a point process on a plane and therefore relies on a different shade of exchangeability (exchangeability for point processes, indeed). It generalises and extends the graphon approach, allows for power-law degree distributions and is projective.

In this work we connect the literature of graphex processes and spatial networks to present a novel methodology to describe sparse spatial random graphs. The model inherits all the desirable properties of the original graphex proposal of Caron and Fox (2017): projectivity, sparsity, interpretability of the parameters and uncertainty quantification. The sociability variables, already present in the base model, ensure scale-free degree distributions for low degree nodes. The addition of a latent spatial component enriches the description of the connections, drawing insights from the topology of the space. Having both sociabilities and locations allows to disentangle the different contributions to edge connections: the connections of a node can be explained both by its sociability and its position in the metric space.

The paper is structured as follows. In section 2 we present the model and its specific construction. Section 3 proves theorems on the asymptotic behaviour of the process, in particular regarding degree distributions, sparsity and clustering. In section 4 we introduce an efficient way to sample from the process and in section 5 we propose a way to perform approximate posterior inference of the variables and parameters. Section 6 discusses the performance of the model with experiment on simulated data. Finally, in section 7 we compare our proposal with different spatial models.

Notation

Consider the stochastic processes $(X_\alpha)_{\alpha\geq 0}, (Y_\alpha)_{\alpha\geq 0}$, $\alpha \in \mathbb{R}_+$, defined on the same probability space and such that $\lim_{\alpha\to\infty} X_\alpha = \lim_{\alpha\to\infty} Y_\alpha = \infty$. We will use $X_\alpha = o(Y_\alpha)$ if and only if $\lim_{\alpha\to\infty} X_\alpha/Y_\alpha = 0$ almost surely, $X_\alpha = O(Y_\alpha)$ if and only if $\limsup_{\alpha\to\infty} X_\alpha/Y_\alpha < \infty$ almost surely, and $X_\alpha = \Theta(Y_\alpha)$ if and only if $X_\alpha = O(Y_\alpha)$ and $Y_\alpha = O(X_\alpha)$ almost surely.

2 The model

In order to introduce our network model we refer to the similar setting introduced in Caron and Fox (2017), Veitch and Roy (2015) and Borgs et al. (2016). Given a sample space F from a probability space $(F, \mathcal{F}, \tilde{F})$, we represent a random graph G as a point process Z defined on $\mathbb{R}_+^2 \times F^2$:

$$Z = \sum_i \sum_j z_{ij} \delta_{(\theta_i, \theta_j, x_i, x_j)} \tag{1}$$

where $i, j \in \mathbb{N}$ are the nodes of the graph and z_{ij} is a binary random variable taking value 1 if nodes i and j are connected by an edge and 0 otherwise. $\theta_i \in \mathbb{R}_+$ is the label of node i and in section 3 we will see how it can be conveniently interpreted as the time of appearance of the node. The variable $x_i \in F$ represents an observable or unobservable variable associated with node i, which we assume to influence the probability of connection between nodes. In this work we will think of x as representing a real or latent location, but it could likewise have different meanings.

We place ourself in the setting introduced by Caron and Fox (2017) by assuming Z to be jointly exchangeable with respect to the labels θ:

$$Z(A_i \times A_j \times \cdot \times \cdot) \stackrel{d}{=} Z(A_{\pi(i)} \times A_{\pi(j)} \times \cdot \times \cdot) \tag{2}$$

for any set $A_i = [(i-1)h, ih), h > 0, i, j \in \mathbb{N}$ and any permutation π of $\mathbb{N}$.

In order to specify a model for $(z_{ij})_{ij}$, we further assign to each node i a variable ϑ_i drawn from a unit-rate Poisson process on a feature space S. S belongs to a σ-finite measure space $(S, \mathcal{S}, \lambda)$, with λ the Lebesgue measure. This also connects our setting to the broader framework introduced in Borgs et al. (2016). If we take S to be the product space $S := \mathbb{R}_+ \times F$, we can define $\vartheta_i := (u_i, x_i)$. Consider the measurable and symmetric function $W : S \times S \to [0, 1]$, known as generalised graphon function. z_{ij} are then conditionally defined as

$$z_{ij} | (\theta_k, \vartheta_k)_k \sim \text{Bernoulli}(W(\vartheta_i, \vartheta_j)). \tag{3}$$

We sample (θ_i, u_i, x_i) from a Poisson process on $\mathbb{R}_+^2 \times F$ with intensity $\lambda(d\theta)\lambda(du)\tilde{F}(dx)$. For background material on point processes we refer to Kingman (1993) and Daley and Vere-Jones (2008). Note that, as proved in Veitch and Roy (2015) and anticipated in section 1, when we restrict the graphon function to a bounded support in the first two coordinates we obtain the original graphon framework that describes only dense or empty graphs. Therefore, we will look at the unbounded support case that yields both sparse and dense graphs.

2.1 Choice of the generalised graphon

We present here the specific generalised graphon function on which we will focus for the rest of the paper. This type of link function is common in network models (see, for example, Aldous

(1997) and Norros and Reittu (2006)) as it implies many nice features. The relation with the variable u will translate in the shape of the degree distribution (see section 3), and the dependence of W on the distance between locations is justified by empirical evidence (Bianconi et al. (2009)). This choice also helps in terms of computational tractability by allowing for an easy switch to the multigraph case (see section 2.3). We consider the feature space $S = \mathbb{R}_+ \times \mathbb{R}^d$ and we fix the generalised graphon function to:

$$\begin{aligned} W((u_i, x_i), (u_j, x_j)) &= 1 - \exp\left(-\frac{2\bar{\rho}^{-1}(u_i)\bar{\rho}^{-1}(u_j)}{(1 + \|x_i - x_j\|)^\gamma}\right) \quad i \neq j \\ W((u_i, x_i), (u_i, x_i)) &= 1 - \exp\left(-\bar{\rho}^{-1}(u_i)^2\right) \end{aligned} \tag{4}$$

yielding the following generative model

$$\begin{aligned} z_{ij}|(\theta_k, u_k, x_k)_{k\geq 1} &\sim \text{Bernoulli}\left(1 - \exp\left(-\frac{2\bar{\rho}^{-1}(u_i)\bar{\rho}^{-1}(u_j)}{(1 + \|x_i - x_j\|)^\gamma}\right)\right) \quad i \neq j \\ z_{ii}|(\theta_k, u_k, x_k)_{k\geq 1} &\sim \text{Bernoulli}\left(1 - \exp\left(-\bar{\rho}^{-1}(u_i)^2\right)\right) \end{aligned} \tag{5}$$

where $\|\cdot\|$ is a norm over $\mathbb{R}^d$ and $\gamma \geq 0$ is the real-valued parameter that tunes the influence of the distance between vertices. Note that for $\gamma = 0$ we revert to the base model of Caron and Fox (2017). ρ is a Lévy measure on $(0, \infty)$, which is $\sigma-$finite and required to be such that $\int_{\mathbb{R}_+} \min(1, w)\rho(dw) < \infty$. $\bar{\rho}$ is the tail Lévy intensity of ρ, which is defined as

$$\bar{\rho}(y) = \int_y^\infty \rho(dw).$$

For the sake of simplicity, we operate a change of variables defining $w_i := \bar{\rho}^{-1}(u_i)$. w therefore belongs to $\mathbb{R}_+$ and the model described by eq. (5) can be rewritten as

$$\begin{aligned} z_{ij}|(\theta_k, w_k, x_k)_{k\geq 1} &\sim \text{Bernoulli}\left(1 - \exp\left(-\frac{2w_i w_j}{(1 + \|x_i - x_j\|)^\gamma}\right)\right) \quad i \neq j \\ z_{ii}|(\theta_k, w_k, x_k)_{k\geq 1} &\sim \text{Bernoulli}\left(1 - \exp\left(-w_i^2\right)\right). \end{aligned} \tag{6}$$

The variables $(w_i, x_i)_{i\geq 1}$ make our model heterogeneous. We interpret the variables w as the sociability weights of the nodes: a very sociable node will display many connections since the link probability function is increasing in w. At the same time, the generalised graphon in eq. (4) is decreasing as a function of the distance between the variables x of two nodes, reflecting the reasonable assumption that similarity in x implies a higher probability of connection. In this new formulation, we sample $(\theta_i, w_i, x_i)_{i\geq 1}$ from a Poisson process with intensity measure $\lambda(d\theta)\rho(dw)\lambda(dx)$. Under this specification, the generative model for $(\theta_i, w_i, x_i)_{i\geq 1}$ is equivalently described through the atomic random measure

$$A = \sum_{i\geq 1} w_i \delta_{(\theta_i, x_i)} \tag{7}$$

which we take to be distributed as a homogeneous completely random measure without deterministic component and with stationary increments, with intensity measure $\lambda(d\theta)\rho(dw)\lambda(dx)$. For background material on completely random measures, we refer to Kingman (1967), Kingman (1993) and Daley and Vere-Jones (2008).

2.2 Choice of the Lévy intensity

The last step to specify our model is the choice of the Lévy measure ρ. This choice will reveal itself crucial to determine most of the theoretical graph properties we wish to achieve. Furthermore, it will allow us to perform efficient simulations, have an approximate posterior inference and interpretable parameters.

An important characteristic of CRMs is their activity, which is a measure of the jump part of the Poisson process and can be infinite or finite. The CRM of eq. (7) has infinite activity when ρ is such that

$$\int_0^\infty \rho(dw) = \infty.$$

In our model, this reflects the fact that the number of potential nodes θ_i, nodes that might or might not be involved in a connection, are in infinite number in every compact subset of $\mathbb{R}_+^2 \times F^2$. As we prove in theorem 3.2, this characteristic is what induces sparsity in the graph. By contrast, if we take a CRM that has finite activity, i.e. with $\int_0^\infty \rho(dw) < \infty$, the resulting graph will be dense.

We already mentioned in section 1 how scale-free degree distributions are very often sought as a desirable goal of network models. After a couple of decades of enthusiasm, though, the debate around these emprical observations and their fit to power-law distributions has sparkled again. Broido and Clauset (2019) showed that under a specific definition of power-law the majority of the degree distributions of a large pool of networks (more than 900) did not pass their goodness-of-fit test. The core of the debate lies around the definition of power-law: in the most restrictive case (considered by Broido and Clauset (2019)), a degree distribution is defined as a power-law when the probability of a node of having degree k is exactly $ck^{-\alpha}$, with c and α positive real numbers and k greater than a fixed threshold $k_{\min}$. Real data, though, have such a complex and noisy nature that this definition is often too strict to be satisfied. As shown in Voitalov et al. (2019), by allowing some slack around the purest power-law definition many more datasets are not rejected as scale-free. This more flexible definition of power-law belongs to the world of regularly varying functions, which we introduce here and can be studied in depth in Bingham et al. (1987). For a detailed account on its use in the random graphs literature, see Van Der Hofstad (2016).

Definition 2.1 (Bingham et al. (1987), pages 6 and 18)**.** *A strictly positive function f on $(0, \infty)$ is said to be **slowly varying at infinity** if $\lim_{y\to\infty} f(ty)/f(y) = 1$ for any $t > 0$. f is **regularly varying at infinity** with exponent $\alpha \in \mathbb{R}$ if it can be written as*

$$f(y) = y^\alpha \ell(y),$$

*with $\ell(y)$ slowly varying at infinity. Finally, f is **regularly varying at 0** with exponent α if $f(1/y)$ is regularly varying at infinity with exponent α, or equivalently*

$$f(y) = y^{-\alpha} \ell(1/y).$$

Intuitively, with this definition we are allowing a power-law behaviour $y^{-\alpha}$ to be smudged by a slowly varying function $\ell(y)$, which does not change too quickly at infinity. Examples of slowly varying functions are functions converging to strictly positive constants or powers of logarithms. Following Van Der Hofstad (2016) (equation 1.4.9), we will place ourselves in this

stream, and define a random variable X to have power-law distribution when its complementaty cumulative distribution function is regularly varying at infinity:

$$\mathbb{P}(X > x) = \ell(x)x^{-\alpha}. \tag{8}$$

To obtain a scale-free degree distribution it is common to impose the same distribution to the sociability weights w (see, for example, Norros and Reittu (2006), Chung and Lu (2002)). Caron et al. (2020) studied the implication of a regularly varying tail Lévy intensity $\bar{\rho}(y)$ on the degree distribution. From now on we place ourselves in their framework and extend their findings to our model, proving asymptotic results on degree distribution, sparsity and clustering. In particular, this means that we will require the tail Lévy measure to be regularly varying at 0. We will outline here two proposals that satisfy this characteristic.

First we consider the generalised gamma process, which has been studied in Hougaard (1986), Aalen (1992), Lee and Whitmore (1993), Brix (1999) and whose role as CRM in Bayesian nonparametrics has been advanced in James (2002), Lijoi and Prünster (2003), Lijoi et al. (2007). Its intensity measure is

$$\rho(dw) = \frac{1}{\Gamma(1-\sigma)} w^{-1-\sigma} e^{-cw} dw \tag{9}$$

with $(\sigma, c) \in (-\infty, 0] \times (0, \infty)$ or $(\sigma, c) \in (0, 1) \times [0, \infty)$. The tail Lévy intensity as y tends to 0 satisfies

$$\bar{\rho}(y) \sim \begin{cases} \frac{1}{\Gamma(1-\sigma)\sigma} y^{-\sigma} & \text{for } \sigma > 0 \\ \log(1/y) & \text{for } \sigma = 0 \\ -c^{\sigma}/\sigma & \text{for } \sigma < 0 \end{cases} \tag{10}$$

which makes it regularly varying at 0 with exponent $\alpha = \max\{0, \sigma\}$. Due to the exponential term in eq. (9), the tail intensity is not regularly varying at infinity. This measure induces both the infinite and finite activity regimes: when $\sigma < 0$ the associated CRM has finite activity, while for $\sigma \geq 0$ it has infinite activity. To achieve sparsity, therefore, we will prefer the latter regime. σ is also pivotal in tuning the exponent of the degree distribution, as we will see in section 3.

When we plug this Lévy measure in the model specified by eq. (5) and γ is strictly positive, then it needs to be such that $\gamma > 1/\min(1, c)$ to assure the degrees to be almost surely bounded, as proved in Deijfen et al. (2013).

In fig. 1 we show three samples from model in eq. (6) with weights sampled from a GG and locations sampled uniformly at random in $[0, 10]^2$. Each network corresponds to a different value of γ in $\{0, 1.5, 3\}$ and the other parameters are selected such that the networks have a comparable number of edges. From the layout of the networks we see how an increase in γ penalises long connections. This is also illustrated in the histograms showing the distribution of the pairwise distances between connected nodes, which show how the mass of the distribution shifts towards the left as the distances gain importance.

2.3 Multigraph representation

One of the reasons to choose the graphon of eq. (4) is because of a handy latent variable representation which induces a directed multigraph and will be useful for simulation purposes, as detailed in section 4. A multigraph is a graph that admits $m_{ij} \in \mathbb{N}$ edges with direction from node i to node j. In our model of eq. (6) this is achieved by setting

$$m_{ij} | (\theta_i, w_i, x_i)_{i \geq 1} \sim \text{Poisson}\left(\frac{w_i w_j}{(1 + \|x_i - x_j\|)^{\gamma}} \right), \quad i \neq j \tag{11}$$

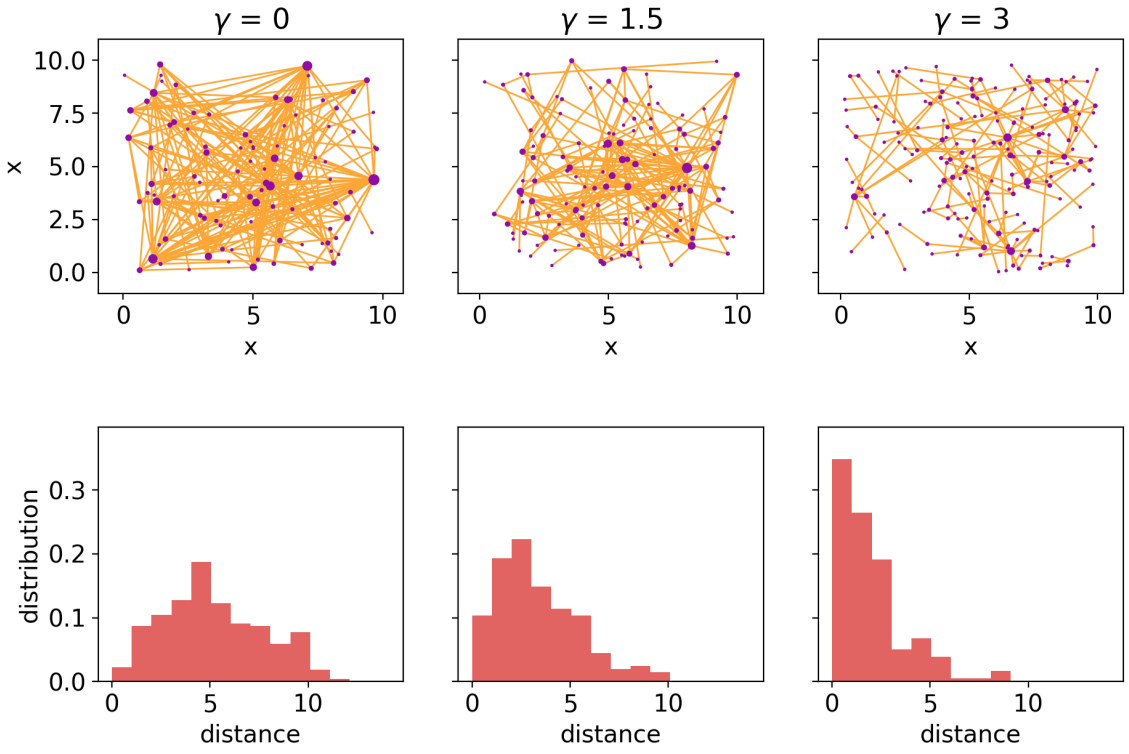


Figure 1: Three samples from model eq. (5), for varying $\gamma \in \{0, 1.5, 3\}$. The Lévy measure is the GG with hyperparameters $c = 1, \sigma = 0.2$ and locations are sampled uniformly at random in $[0, 10]^2$. t, parameter introduced in section 3 to sample finite graphs, takes respectively the values $20, 50$ and 90 so that the number of edges is comparable across plots (around 220). The number of nodes is, from left to right, $98, 128$ and 193. The top row represents the network in the space $[0, 10]^2$, with the node sizes proportional to their degrees. On the bottom row, the respective histograms of the pairwise distances between connected nodes.

$$m_{ii}|(\theta_i, w_i, x_i)_{i\geq 1} \sim \text{Poisson}\left(w_i^2\right)$$

from which we recover the original model by simply checking whether $m_{ij} + m_{ji}$ is positive or null: $z_{ij} = \mathbb{1}_{m_{ij}+m_{ji}>0}$. We can equivalently represent the multigraph with a Poisson random measure

$$M = \sum_{i\neq j} m_{ij}\delta_{(\theta_i,\theta_j,x_i,x_j)} + \sum_i m_{ii}\delta_{(\theta_i,\theta_i,x_i,x_i)} \tag{12}$$

with mean measure

$$\mu = \sum_{i\neq j} \frac{w_i w_j}{(1+\|x_i - x_j\|)^\gamma}\delta_{(\theta_i,\theta_j,x_i,x_j)} + \sum_i w_i^2\delta_{(\theta_i,\theta_i,x_i,x_i)}. \tag{13}$$

3 Time asymptotic properties

The nature of the point process of eq. (1) prescribes the number of nodes and edges of a graph realisation to be almost surely infinite. To obtain graphs whose number of edges is finite, we consider a truncated domain for the label $\theta \in [0, t], t > 0$. As anticipated in the introduction, it comes in handy to think of θ as time of arrival of the node: when we restrict the point process to $\theta \leq t$ we mean that nodes whose arrival time exceeds t will not be considered. In this section we will focus on the behaviour of the graphex process as the threshold t goes to infinity: the intuitive idea is that, as time goes by, new nodes appear and connect to the already existing nodes and we can observe the asymptotic behaviour of the process. For the same reason, we restrict the space domain to an interval $[0, x_{\max}], x_{\max} > 0$. Even in the truncated process, though, the number of potential nodes and connections is infinite when the activity of the Lévy measure is infinite, as explained in the last section 2.2. To have finite graphs we need to take a step further and limit ourselves to active nodes, i.e. those that have positive degree. Note that we study the process as time goes to infinity, which is not the usual type of asymptotics in space ($x_{\max}$ going to infinity) that spatial networks model study (see, for example Deijfen et al. (2013), Deprez and Wüthrich (2018), Dalmau and Salvi (2019)). Our approach is motivated by the similitude with Caron et al. (2020) and Veitch and Roy (2015), who studied the asymptotics for the base graphex framework and from which we extend our results. The spatial asymptotics remains a very interesting open problem to explore in the future, on which we make some conjectures in section 7.

Let us first introduce the statistics of interest. To indicate the truncation in time, we will use t as subscript for the various processes. For simplicity of notation, we will omit the indication of the truncation in space. The truncated version of the process in eq. (1) becomes

$$Z_t = \sum_{ij} z_{ij}\delta_{(\theta_i,\theta_j,x_i,x_j)}\mathbb{1}_{\theta_i\leq t}\mathbb{1}_{\theta_j\leq t}. \tag{14}$$

The degree of node i is the number of nodes connected to i

$$D_{t,i} = \sum_j z_{ij}\mathbb{1}_{\theta_i\leq t}\mathbb{1}_{\theta_j\leq t}. \tag{15}$$

The number of active nodes is the number of nodes with positive degree

$$N_t = \sum_i \mathbb{1}_{D_{t,i}\geq 1}\mathbb{1}_{\theta_i\leq t}, \tag{16}$$

and the number of nodes with degree j is

$$N_{t,j} = \sum_i \mathbb{1}_{D_{t,i}=j} \mathbb{1}_{\theta_i \leq t}. \tag{17}$$

The total number of edges in the indirected graph is

$$N_t^{(e)} = \sum_i z_{ii} \mathbb{1}_{\theta_i \leq t} + \sum_{j>i} z_{ij} \mathbb{1}_{\theta_i \leq t} \mathbb{1}_{\theta_j \leq t}. \tag{18}$$

We can finally state formally the definition of dense and sparse graphs.

Definition 3.1. *A family of graphs $(G_t)_{t\geq 0}$ associated with the point process in eq. (14) with N_t nodes and $N_t^{(e)}$ edges is **dense** if*

$$N_t^{(e)} = \Theta(N_t^2) \text{ as } t \to \infty, \tag{19}$$

*and **sparse** if*

$$N_t^{(e)} = o(N_t^2) \text{ as } t \to \infty. \tag{20}$$

Recall that in a simple graph (i.e. with no multiple edges and self loops) with N_t vertices, the maximum number of possible edges is $N_t(N_t - 1)/2 = \Theta(N_t^2)$, justifying the previous definition. At this point, we can adapt one of the most important results of Caron and Fox (2017) about sparsity and activity of the Lévy measure to our setting.

Theorem 3.2. *Consider the point process Z_t of eq. (14) representing the model specified in eq. (6), with N_t and $N_t^{(e)}$ the associated number of nodes and edges. Consider a Lévy measure ρ such that $\int_0^\infty w\rho(dw) < \infty$. If the CRM has finite activity, i.e. $\int_0^\infty \rho(dw) < \infty$, then the graph is dense. If the CRM has infinite activity, i.e. $\int_0^\infty \rho(dw) = \infty$, then the graph is sparse.*

Proof. The proof is presented in appendix 9.1. □

As highlighted in the introduction, another important feature of real-world networks is clustering, a measure of transitivity in graphs. Clustering coefficients generally measure how strongly a group of nodes is interconnected. The strength is usually measured by reshuffling the edges of the graph uniformly at random and measuring the density of the edges in the considered group of nodes before and after the reshuffling. A density that decreases after this procedure indicates a positive clustering among those nodes, while if we observe an increase then the clustering is negative. Depending on the zoom level you are using, different clustering measures have been proposed. In this work we will focus on the asymptotic behaviour of the global and average local clustering coefficients.

Definition 3.3. *The number of triangles in which node i is involved is*

$$T_{t,i} = \frac{1}{2} \sum_{j\neq i, k\neq i, k\neq j} z_{ij} z_{ik} z_{jk} \mathbb{1}_{\theta_i \leq t} \mathbb{1}_{\theta_j \leq t} \mathbb{1}_{\theta_k \leq t}.$$

*The **global clustering coefficient** is the ratio of the number of triangles over the number of triplets in the graph:*

$$C_t^{(g)} = \frac{3 \times \textit{ number of triangles}}{\textit{number of open and closed triplets}} = \frac{\sum_i T_{t,i}}{\sum_{i\neq j\neq k} z_{ij} z_{ik} \mathbb{1}_{\theta_i \leq t} \mathbb{1}_{\theta_j \leq t} \mathbb{1}_{\theta_k \leq t}}. \tag{21}$$

The local clustering coefficients are the analogous of the global coefficient when restricted to nodes of degree j, for $j \geq 2$:

$$C_{t,j}^{(l)} = \frac{2}{j(j-1)N_{t,j}} \sum_i T_{t,i} \mathbb{1}_{D_{t,i}=j}.$$

*The **average local clustering coefficient** $C_t^{(l)}$ is the mean of the $C_{t,j}^{(l)}$:*

$$C_t^{(l)} = \frac{1}{N_t - N_{t,1}} \sum_{j \geq 2} N_{t,j} C_{t,j}^{(l)}. \tag{22}$$

We present now the results about the asymptotic limits of the number of nodes, edges, fraction of nodes with a certain degree and clustering coefficients. Theorem 3.4 deals with sparsity and degree distribution and requires the regular variation assumption. Theorem 3.5 is divided into two parts: the former, about the global clustering coefficient, does not require assumptions on ρ, while the result about the local clustering requires regular variation and some integrability constraints. In remark 1 we specialise the obtained results to the Lévy measure proposed in eq. (9). In remark 2 we talk about possible generalisations of our results to a broader class of generalised graphon functions. Finally, we show some plots to illustrate the asymptotic behaviours on simulated data. Theorems 3.4 and 3.5 are applications of broader results respectively found in Caron et al. (2020) (remark 26) and Borgs et al. (2019) (Proposition 56).

Theorem 3.4. *Consider a random graph family $(G_t)_{t \geq 0}$ associated with point process Z_t of eq. (14) and the model of eq. (5). Let the associated Lévy measure ρ have a tail intensity regularly varying at 0 with exponent $\alpha < 1$.*

For $\alpha \in (0, 1)$ the graph is sparse and the number of edges and nodes obey to the following relation:

$$N_t^{(e)} = \Theta(N_t^{1+\alpha}).$$

For $\alpha = 0$ we observe a sparse, almost dense behaviour

$$N_t^{(e)} = \Theta\left(\frac{N_t^2}{\log(N_t)^2}\right),$$

while for $\alpha < 0$ the graph is dense:

$$N_t^{(e)} = \Theta(N_t^2). \tag{23}$$

As for the degree distribution, for a strictly positive α the fraction of nodes with positive degree j is such that

$$\frac{C_1}{C_2} \frac{\alpha \Gamma(j-\alpha)}{j! \Gamma(1-\alpha)} \leq \lim_{t \to \infty} \frac{N_{t,j}}{N_t} \leq \frac{C_2}{C_1} \frac{\alpha \Gamma(j-\alpha)}{j! \Gamma(1-\alpha)}$$

for some positive constants C_1, C_2. Note that, as j goes to infinity, we have $\alpha\Gamma(j-\alpha)/(j!\Gamma(1-\alpha))$ converging to $\alpha j^{-1-\alpha}/\Gamma(1-\alpha)$ and hence a power-law behaviour with exponent $1+\alpha$ for large degrees. For $\alpha \leq 0$, the fraction of nodes with positive degree j converges to 0 almost surely:

$$\frac{N_{t,j}}{N_t} \overset{a.s.}{\to} 0.$$

Proof. The proof of these results can be found in appendix 9.1. □

Theorem 3.5. *Consider a random graph family $(G_t)_{t\geq 0}$ associated with the point process Z_t of eq. (14) and the model in eq. (5). Let*

$$\int (1 - e^{-\frac{2w_iw_j}{(1+\|x_i-x_j\|)^\gamma}})(1 - e^{-\frac{2w_iw_k}{(1+\|x_i-x_k\|)^\gamma}})\rho(dw_i)\rho(dw_j)\rho(dw_k)dx_idx_jdx_k < \infty. \tag{24}$$

Then, we have that the global clustering coefficient has an almost sure limit as t tends to infinity:

$$C_t^{(g)} \xrightarrow{a.s.} \frac{\int (1 - e^{-\frac{2w_iw_j}{(1+\|x_i-x_j\|)^\gamma}})(1 - e^{-\frac{2w_iw_k}{(1+\|x_i-x_k\|)^\gamma}})(1 - e^{-\frac{2w_kw_j}{(1+\|x_k-x_j\|)^\gamma}})\rho(dw_i)\rho(dw_j)\rho(dw_k)d\mathbf{x}}{\int (1 - e^{-\frac{2w_iw_j}{(1+\|x_i-x_j\|)^\gamma}})(1 - e^{-\frac{2w_iw_k}{(1+\|x_i-x_k\|)^\gamma}})\rho(dw_i)\rho(dw_j)\rho(dw_k)d\mathbf{x}}. \tag{25}$$

In the hypothesis of ρ regularly varying at 0 with exponent $\alpha \in (0, 1)$ and when

$$b := \frac{\int \frac{w_jw_k}{(1+\|x_i-x_j\|)^\gamma(1+\|x_i-x_k\|)^\gamma}(1 - e^{-\frac{2w_jw_k}{(1+\|x_j-x_k\|)^\gamma}})\rho(dw_j)\rho(dw_k)dx_idx_jdx_k}{(\int \frac{w_j}{(1+\|x_i-x_j\|)^\gamma}\rho(dw_j)dx_idx_j)^2} \tag{26}$$

is finite and belongs to the interval $(0, 1]$, then the local clustering coefficient has limit in probability for t that tends to infinity:

$$C_t^{(l)} \xrightarrow{P} b.$$

Proof. The proof can be found in appendix 9.1. □

Remark 1 (GG measure)**.** *It is easy to prove directly that theorem 3.5 holds for the GG measure. Consider eq. (24) and observe that*

$$\begin{aligned}&\int (1 - e^{-\frac{2w_iw_j}{(1+\|x_i-x_j\|)^\gamma}})(1 - e^{-\frac{2w_iw_j}{(1+\|x_i-x_k\|)^\gamma}})\rho(dw_i)\rho(dw_j)\rho(dw_k)dx_idx_jdx_k\\ &\leq x_{\max}^3 \int \left[\int (1 - e^{-2w_iw_j})\rho(dw_j)\right]^2 \rho(dw_i).\end{aligned} \tag{27}$$

Define

$$\psi(w) := \int_0^\infty (1 - e^{-2ww_j})\rho(dw_j),$$

substitute the Lévy measure of the GG, eq. (9), and use the approximation $1 - e^{-x} \sim x$ as x tends to 0:

$$\begin{aligned}\psi(w) &= \int_0^1 (1 - e^{-2ww_j})w_j^{-1-\sigma}e^{-cw_j}dw_j + \int_1^\infty (1 - e^{-2ww_j})w_j^{-1-\sigma}e^{-cw_j}dw_j\\ &\sim w\int_0^1 w_j^{-\sigma}e^{-cw_j}dw_j + \int_1^\infty (1 - e^{-2ww_j})w_j^{-1-\sigma}e^{-cw_j}dw_j.\end{aligned}$$

By definition of eq. (9), $\sigma < 1$ and therefore the first integral converges, while the second one is always finite due to the exponential decay. Therefore, $\psi(w) \sim w$. Therefore, eq. (27) is proportional to

$$x_{\max}^3 \int_0^\infty w_i^{2-1-\sigma}e^{-cw_i}dw_i.$$

Similarly, as $\sigma < 1$ the integral is convergent. The hypothesis for convergence of the global clustering coefficient holds for weights generated under the GG process. With very similar computations it is possible to show that eq. (26) is finite for the GG process.

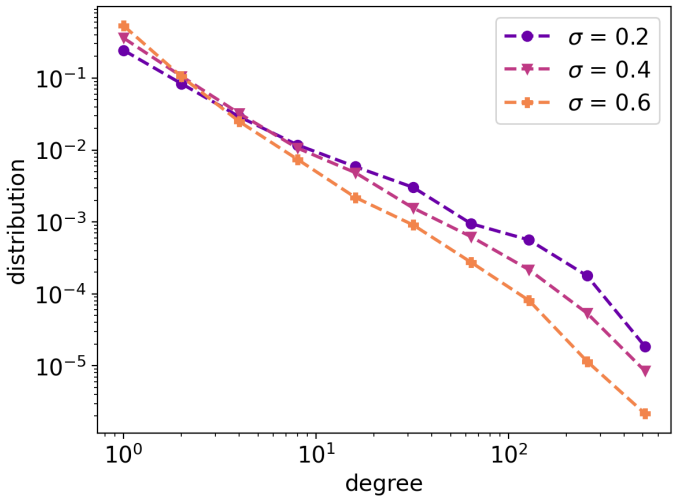


Figure 2: Degree distribution (in logarithmic scale) of graphs generated under model eq. (6) with sociabilities sample from a GG eq. (9) with varying $\sigma \in \{0.2, 0.4, 0.6\}$. Locations are sampled uniformly at random in $[0, 10]$. The truncation level of the GG is $\epsilon = 10^{-4}$ and remaining parameters are fixed to $t = 150, c = 0.1, \gamma = 1$.

In fig. 2 we can visualise different power-law behaviours of the degree distribution as a function of the regular variation exponent σ of the GG. As proved in theorems 3.4 and 3.2, this coefficients tunes both the sparsity level and the slope of the power-law degree distribution.

In fig. 3 we observe the clustering coefficients' asymptotic behaviour.

It is interesting to note how even for $\gamma = 0$ the model is able to achieve positive clustering. This is in stark contrast with usual inhomogeneous non spatial models, which lack in clustering and usually require the addition of a space component to achieve it (see, for example, Deijfen et al. (2013) and Bringmann et al. (2019)). The motivation lies in the fact that our asymptotic studies are in time and not in space, in contrast with the usual spatial approach. This is an interesting difference among the two approaches, which makes it challenging to compare them directly but suggests possible future work directions.

Remark 2 (Choice of the generalised graphon function)**.** *The results of theorems 3.4 and 3.5 do not hold solely for a generalised graphon function of the type of eq. (6). In fact, as illustrated in Caron et al. (2020), the framework is flexible and can be used for any function W that can be factorised as a the following product:*

$$W(\vartheta_i, \vartheta_j) = \eta(u_i, u_j)\omega(x_i, x_j)$$

where $\eta : \mathbb{R}_+ \times \mathbb{R}_+ \to [0, 1]$ is the component capturing the sparsity and $\omega : F \times F \to [0, 1]$ deals with the local structure (for example distances or communities). Under some regular variation constraints for the function $\mu_\eta(u) = \int \eta(u, u')du'$ and bounded behaviour of $\nu_\eta(u, u') = \int \eta(u, \tilde{u})\eta(u', \tilde{u})$ illustrated in Caron et al. (2020) the previous theorems hold similarly.

The same is true for a more flexible class of generalised graphons. Consider

$$\mu(u, x) := \int W((u, x), (u', x'))du'\tilde{F}(dx')$$

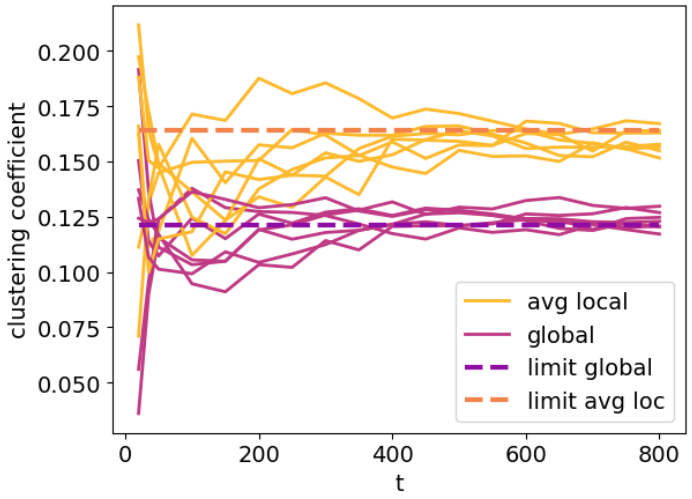


Figure 3: Global and average local clustering coefficients as a function of t for 7 samples of graphs with weights sampled from a generalised gamma process eq. (9) with $\sigma = 0.2, c = 2$ and truncation level $\epsilon = 10^{-4}$, and locations drawn uniformly at random in $[0, 1]$. γ is fixed to 1. The dashed lines represent the asymptotic limits derived in theorem 3.5.

and assume that for each fixed x there exist $u_0(x) > 0$ such that for $u > u_0$

$$C\mu_{\tilde{\eta}}(u)\mu_{\tilde{\omega}}(x) \leq \mu(u, x) \leq C'\mu_{\tilde{\eta}}(u)\mu_{\tilde{\omega}}(x) \tag{28}$$

for some positive functions $\tilde{\eta} : \mathbb{R}_+ \to \mathbb{R}_+$ and $\tilde{\omega} : F \to \mathbb{R}_+$ and positive constants C, C'. Assume $\mu_{\tilde{\eta}}$ and $\nu_{\tilde{\eta}}$ satisfy Assumptions 1 and 2 of Caron et al. (2020), then the results of the previous theorems can be adapted to this framework. In particular, the condition described in eq. (28) holds for Ws of the form

$$W((u_i, x_i), (u_j, x_j)) = 1 - e^{-\tilde{\eta}(u_i, u_j)\tilde{\omega}(x_i, x_j)}$$

Our model enters into this second framework as

$$\tilde{\eta}(u_i, u_j) = 2\bar{\rho}^{-1}(u_i)\bar{\rho}^{-1}(u_j)$$
$$\tilde{\omega}(x_i, x_j) = \frac{1}{(1 + \|x_i - x_j\|)^{\gamma}}$$

with $\tilde{F}(dx) = dx$ and $x \in \mathbb{R}^d$. The function $\tilde{\omega}$ could be even more general than this, as it can be extended to any function $\tilde{\omega}(x_i, x_j) = g(\|x_i - x_j\|)$ with $g : \mathbb{R}_+ \to \mathbb{R}_+$ and $\|\cdot\|$ norm on F.

4 Sampling algorithms

In this section we discuss the challenge of efficient sampling of graphs from our model. For simplicity, we focus on the case $d = 1$, with d the dimension of the space of x.

We explained at the beginning of section 3 how to obtain a finite graph: we restrict the process in time $\theta \in [0, t]$ and space $x \in [0, x_{\max}]$, and we consider only nodes with positive degree. A priori, though, we do not know which w will be associated with a positive degree and we need to find a way to avoid sampling an infinite number of them, which is the case for infinite activity CRMs. Therefore, we will present ways to simulate a graph that rely on truncation techniques.

In particular, we first offer a naive sampling technique with time complexity $O(N^2)$ (with N the number of nodes) and then, in section 4.1 we propose a way to decrease the complexity of the naive algorithm to $O(N \log^2 N)$.

We mention here that the exact simulation of a CRM with infinite activity is possible for some specific Lévy measures. In such cases (for example, the GG of eq. (9)) we need not revert to truncation and could sample a graph with different techniques. We do not explore this approach here, but it can be derived very similarly to section 5.5.2 of Caron and Fox (2017).

The restricted CRM representing the generative process of $(\theta_i, w_i, x_i)_{i\geq 1}$ is

$$A_{t,x_{\max}} = \sum_{i\geq 1} w_i \delta_{(\theta_i,x_i)} \mathbb{1}_{x_i\in[0,x_{\max}]} \mathbb{1}_{\theta_i\in[0,t]} \tag{29}$$

The corresponding restricted Poisson random measure equivalent to eq. (12) is

$$M_{t,x_{\max}} = \sum_{i,j} m_{ij} \delta_{(\theta_i,\theta_j,x_i,x_j)} \mathbb{1}_{x_i,x_j\in[0,x_{\max}]} \mathbb{1}_{\theta_i,\theta_j\in[0,t]} \tag{30}$$

with intensity

$$\begin{aligned}\mu_{t,x_{\max}} = &\sum_{i\neq j} \frac{w_i w_j}{(1+|x_i-x_j|)^\gamma} \delta_{(\theta_i,\theta_j,x_i,x_j)} \mathbb{1}_{x_i,x_j\in[0,x_{\max}]} \mathbb{1}_{\theta_i,\theta_j\in[0,t]} \\ &+ \sum_i w_i^2 \delta_{(\theta_i,\theta_i,x_i,x_i)} \mathbb{1}_{x_i\in[0,x_{\max}]} \mathbb{1}_{\theta_i\in[0,t]}\end{aligned} \tag{31}$$

In the infinite activity case, a simple approximate approach to simulate the weights w_i, applicable to any Lévy measure, is the inverse Lévy method: after fixing a threshold $\epsilon > 0$, we sample weights from the truncated CRM

$$A^\epsilon_{t,x_{\max}} = \sum_i w_i \delta_{(\theta_i,x_i)} \mathbb{1}_{x_i,x_j\in[0,x_{\max}]} \mathbb{1}_{\theta_i,\theta_j\in[0,t]} \mathbb{1}_{w_i\geq\epsilon}. \tag{32}$$

The random measure that describes the multigraph becomes

$$M^\epsilon_{t,x_{\max}} = \sum_{i,j} m_{ij} \delta_{(\theta_i,\theta_j,x_i,x_j)} \mathbb{1}_{x_i,x_j\in[0,x_{\max}]} \mathbb{1}_{\theta_i,\theta_j\in[0,t]} \mathbb{1}_{w_i,w_j\geq\epsilon}. \tag{33}$$

Hence, for finite activity or truncated infinite activity CRMs we can easily simulate a graph with the naive procedure of drawing edges for every possible pair of nodes. We describe this procedure in algorithm 1.

Algorithm 1 Sampling Algorithm

Input: $x_{\max}, t, \epsilon, \gamma$, the Lévy measure's parameters, empty square matrix G
Output: G

Sample $(w_i)_{i\geq 1}$ from the CRM $A_{t,x_{\max}}$ or from $A^\epsilon_{t,x_{\max}}$. Call N the cardinality.
Sample $(x_i)_{i=1}^N, (\theta_i)_{i=1}^N$ uniformly at random over $[0, x_{\max}]$ and $[0, t]$.
for $i = 1, \ldots, N$ **do**
    Sample $m_{ii} \sim$ Poisson (w_i^2)
    if $m_{ii} > 0$ **then**
        Set $G(i, i) = 1$.
    for $j = i + 1, \ldots, N$ **do**
        Sample $m_{ij} \sim$ Poisson $\left(\frac{2w_i w_j}{(1+|x_i-x_j|)^\gamma}\right)$.
        if $m_{ij} > 0$ **then**
            Set $G(i, j) = G(j, i) = 1$.

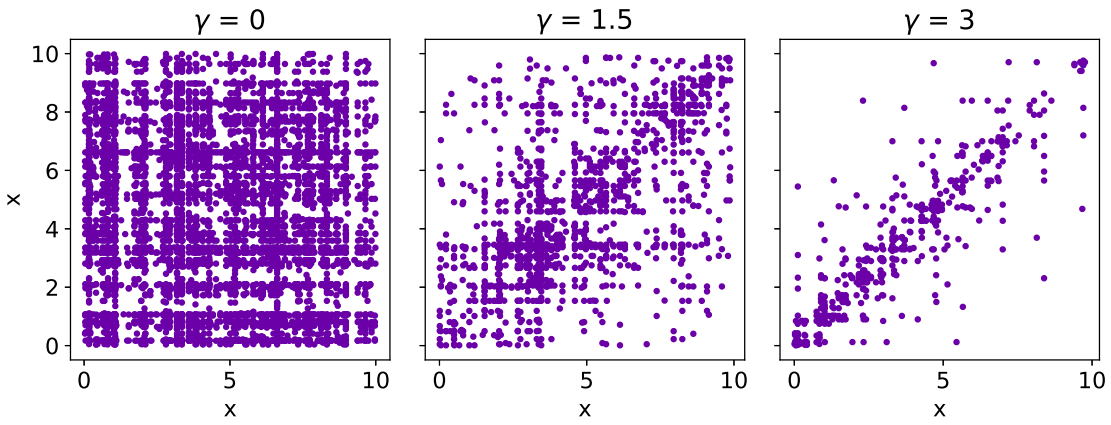


Figure 4: Adjacency matrices of three samples from the sparse spatial model for γ varying in $\{0, 1.5, 3\}$. The sociability weights are sampled from the GG process eq. (9) with $\sigma = 0.2, c = 0.7$ and truncation level $\epsilon = 10^{-4}$, and locations are drawn uniformly at random in $[0, 10]$. The dots represent a connection between nodes with coordinates x_i, x_j.

Since for every node we have to cycle through every possible neighbour, the time complexity of this algorithm is of order $O(N^2)$. In fig. 4 we represent samples from the model represented through the adjacency matrix indexed by the values of x. Figure 4 is particularly important because it leads our way to the next paragraph: for $\gamma > 0$ the majority of connections happen to be along the diagonal, i.e. connecting nodes close to each others in space. The magnitude of the connections, therefore, looks linear in space and gives us hope to find an algorithm that will exploit wisely the spatial information to decrease the complexity.

4.1 Efficient sampling via grids

Knowing that our interest is for graphs whose number of edges is below the squared number of nodes, we would like to find a smart way to compute $w_i w_j/(1 + |x_i - x_j|)^\gamma$ only for pair of nodes that are likely to be connected. Based on the observations made from fig. 4, we partition the space domain into cells to leverage the information that close cells will contain more edges than cells that lie far away in space, as illustrated on the left of fig. 5 where we have divided the domain $[0, 10]$ into 1-dimensional boxed of size 2.5. Formally, we construct a regular grid of K cells B_k of size $\delta = x_{\max}/K$ and we take $K = O(N)$. Taking inspiration from Bringmann et al. (2019), we perform a similar operation for the sociability weights, slicing them into layers. The layers are not all equispaced, as the spatial grid, but have exponential growth and are defined as:

$$V_j := \{i \in \{1, \dots, N\} : \underline{w}_j := \underline{w}_0 2^j \leq w_i < \underline{w}_{j+1} := \underline{w}_0 2^{j+1}\} \quad j = 1, \dots, \overline{J} \tag{34}$$

with $\underline{w}_0 := \min_{i\in\{1,\dots,N\}} w_i$ and $\overline{J} := \log_2(\max_{i\in\{1,\dots,N\}} w_i/w_0)$. For an illustration, see fig. 5.

On a high level, we associate each node with a spatial cell and a sociability layer in order to bound its probability of connection with some constants that are common to all the nodes that belong to that cell and layer. Thinning techniques for Poisson processes are the key that allows us to propose a new algorithm that still samples exactly according to the model of eq. (6)

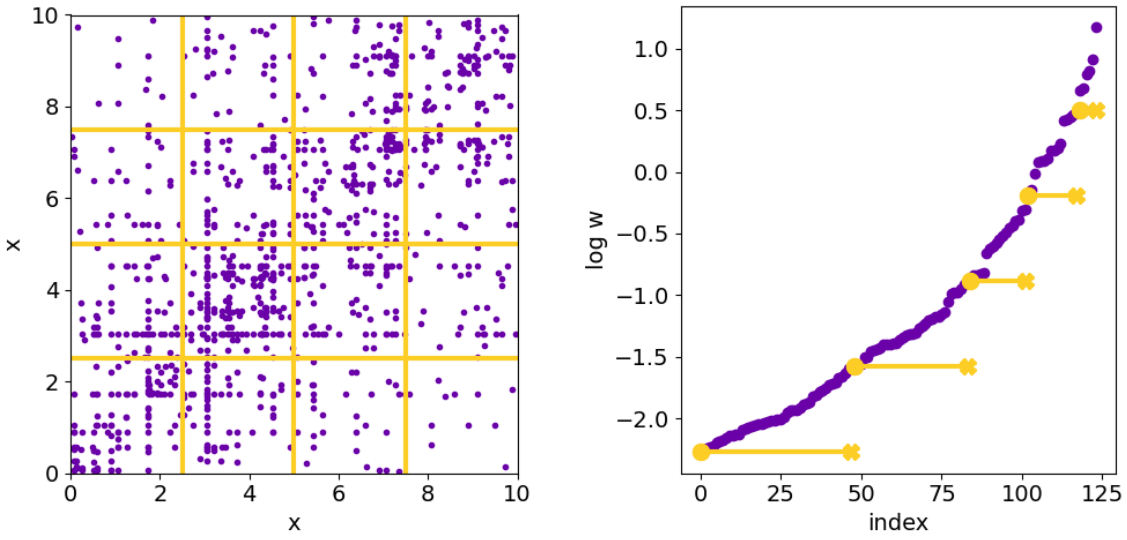


Figure 5: On the left, adjacency matrix plotted according to the nodes' space coordinates and spatial grid of size 2.5. On the right, the logarithm of the sociability weights in increasing order (purple) and layers of the weights (yellow). The graph has been sampled from the sparse spatial model with $\gamma = 1.5$, sociability weights sampled from the GG process with $\sigma = 0.2, c = 0.7$ and truncation level $\epsilon = 10^{-4}$, and locations drawn uniformly at random in $[0, 10]$. Space is partitioned into a grid with size 2.5.

and whose complexity is $O(N \log^2 N)$. For a reference on thinning methods, see Stoyan et al. (2013), and for the details of the algorithm and proof of the complexity we refer to appendix 9.2.

Intuitively, the algorithm works as follows. For nodes in adjacent cells B_k, B_{k+1} we sample the edges according to algorithm 1, while the algorithm changes for cells that are not adjacent. For each cell B_k and each pair of layers V_{j_1}, V_{j_2} we sample an upper bound on the total number of connections whose starting point is a node in cell B_k and layer V_{j_1} and end point is a node in layer V_{j_2}. Then, we sample the cell $B_l, l = k+2, \dots, K$ where the end point of the edge will lie and and eventually we sample the pair of nodes associated to that edge.

Algorithm 2 describes exactly the steps of the procedure. For the sake of simplicity, we focus on the finite activity case. Everything holds in the same way for the infinite activity case with the truncated process and intensity defined in eq. (29) and eq. (32). New notation needs to be defined for algorithm 2. In particular, $p(l; k, K, \gamma, \delta) := 1/(C_{k,K,\gamma,x_{\max}}(1 + (-\delta + \delta l))^{\gamma})$ is the probability mass function of a Zipf distribution on the cells $B_{k+2}, \dots, B_K$ that can contain the end point of a connection starting from B_k, of which $C_{k,K,\gamma,\delta} = \sum_{l=k+2}^{K}(1 - \delta + \delta|k - l|)^{-\gamma}$ is the normalizing constant. $\overline{w}_k^{(j)}$ is the sum of the weights associated to nodes that belong to layer V_j and cell B_k: $\overline{w}_k^{(j)} = \sum_i w_i \mathbb{1}_{i \in V_j} \mathbb{1}_{i \in B_k}$.

Algorithm 2 Sampling Algorithm

Input: $K, \gamma, t, x_{\max}, \epsilon$, the Lévy measure parameters, empty square matrix G
Output: G

Sample $(w_i)_{i\geq 1}$ from the CRM $A_{t,x_{\max}}$ or $A^{\epsilon}_{t,x_{\max}}$. Call N the cardinality.
Sample $(x_i)_{i=1}^{N}, (\theta_i)_{i=1}^{N}$ uniformly at random over $[0, x_{\max}]$ and $[0, t]$.
Create weight layers $V_j, j = 1 \dots, \overline{J}$ as described in eq. (34).
Create the partition $B_1, \dots, B_K$ of equal side length of $[0, x_{\max}]$. Set $\delta = x_{\max}/K$.
Create vector $(C_{k,K,\gamma,\delta})_{k=1}^{K-2}$.
for $j_1 = 1, \dots, \overline{J}$ **do**
    for $l = 1, \dots, K-2$ **do**
        Compute $\overline{w}_l^{(j_1)} = \sum_i w_i \mathbb{1}_{i\in V_{j_1}} \mathbb{1}_{i\in A_l}$.
for $j_1 = 1, \dots, \overline{J}$ **do**
    for $j_2 = 1, \dots, \overline{J}$ **do**
        for $k = 1, \dots, K-2$ **do**
            Sample $m_k^{(j_1,j_2)} \sim \text{Poisson}(2\overline{w}_k^{(j_1)} \underline{w}_{j_2+1} |V_{j_2}| C_{k,K,\gamma,\delta})$
            for $m = 1, \dots, m_k^{(j_1,j_2)}$ **do**
                Sample $u_{km}^{(j_1,j_2)} \sim p(\cdot)$ and set $l := u_{km}^{(j_1,j_2)}$.
                if Uniform$[0,1] < \frac{\overline{w}_l^{j_2}}{\underline{w}_{j_2+1}|V_{j_2}|}$ **then**
                    Sample an edge from pmf $\frac{w_i}{\overline{w}_k^{j_1}} \times \frac{w_j}{\overline{w}_l^{j_2}}$ for $x_i, x_j \in B_k \times B_l$.
                    if Uniform$[0,1] < \frac{(1-\delta+\delta|k-l|)^{\gamma}}{(1+|x_i-x_j|)^{\gamma}}$ **then**
                        Set $G(i,j) = 1$.
        for $k = K-1, K$ **do**
            follow algorithm algorithm 1 for nodes in these cells and layers

5 Posterior inference

In this section we propose an algorithm to perform posterior inference in the case of the model in eq. (6). We observe a set of connections $(z_{ij})_{1\leq i,j\leq N_t}$ among N_t active nodes. We aim at sampling from the posterior

$$p\left((w_i)_{i=1}^{N_t'}, (x_i)_{i=1}^{N_t'}, \phi, t, \gamma | (z_{ij})_{i,j=1}^{N_t}\right). \tag{35}$$

where ϕ is the set of parameters of the Lévy measure ρ and $N_t' > N_t$ represents the unknown cardinality of the full set of nodes, active and inactive. Note that N_t' is infinite for infinite activity CRMs. Observe that this is the inference setting of a latent space model, or space model where x is unknown. If x is treated as a covariate, the posterior becomes

$$p\left((w_i)_{i=1}^{N_t'}, (x_i)_{i=N_t+1}^{N_t'}, \phi, t, \gamma | (z_{ij})_{i,j=1}^{N_t}, (x_i)_{i=1}^{N_t}\right).$$

For our choices of ρ (eq. (9)) the posterior in eq. (35) is not tractable, and we will therefore rely on a Markov chain Monte Carlo (MCMC) approach to sample values approximately distributed from the posterior. Our algorithm will exploit steps from Gibbs samplers, Metropolis-Hastings (MH) and Hamiltonian Monte Carlo (HMC). For references about these methods, see respectively Geman and Geman (1984), Hastings (1970) and Neal (2011).

In the following subsection, we will present the case of the weights from a generalised gamma process of eq. (9) and specify the remaining prior distributions for the hyperparameters, parameters and locations.

5.1 Approximate posterior inference for the generalised gamma process

To overcome the challenge of infinite activity CRMs, we rely on the finite and independent identically distributed approximation of CRMs introduced in Lee (2019) and Lee et al. (2016), a brief summary of which can be found in appendix 9.3. Such approximation of an infinite activity CRM is a finite CRM $\tilde{A}_L$ whose weights are independently and identically distributed from an appropriate density f:

$$\sum_{i=1}^{L} \tilde{w}_i \delta_{(\theta_i, x_i)}, \quad \tilde{w}_i \overset{iid}{\sim} f \tag{36}$$

such that $\sum_{i=1}^{L} \tilde{w}_i \delta_{(\theta_i, x_i)}$ converges in distribution to the infinite dimensional CRM A of eq. (7).

Consider $\rho(dw)$ the Lévy measure of a generalised gamma process with parameters σ and c, $\sigma \in (0,1), c > 1$, as introduced in eq. (9). As presented in Lee (2019), a finite independent and identically distributed representation of the generalised gamma process is offered by the exponentially-tilted BFRY distribution with parameters (σ, ζ, c), whose density function is

$$f_\zeta(d\tilde{w}) = \frac{\sigma \tilde{w}^{-1-\sigma} e^{-c\tilde{w}} (1 - e^{-\zeta \tilde{w}})}{\Gamma(1-\sigma)((\zeta + c)^\sigma - c^\sigma)} d\tilde{w} \tag{37}$$

where $\zeta = (L\sigma/t)^{1/\sigma}$.

Our new target is then the following approximation of the posterior in eq. (35), truncated at $L > N_t$:

$$p\left((\tilde{w}_i)_{i=1}^L, (x_i)_{i=1}^L, \phi, t, \gamma | (z_{ij})_{i,j=1}^L\right). \tag{38}$$

with $\phi = \{\sigma, c\}$. Note that $z_{ij} = 0$ for $i, j = N_t{+}1, \ldots, L$, $i = 1, \ldots, N_t$ and $j = N_t{+}1, \ldots, L$, $j = 1, \ldots, N_t$ and $i = N_{t+1}, \ldots, L$.

Prior distributions and MCMC algorithm

We present here the general structure of the MCMC procedures to sample from the posterior presented in eq. (38) in the case of the GG process prior for weights, which we approximate with the distribution of eq. (37). For details on the proposals and computations we refer to appendix 9.4. For simplicity, we define $p_{ij} := (1 + |x_i - x_j|)^{-\gamma}$.

Following Caron and Fox (2017), we place a Gamma prior on t and improper priors on σ and c:

$$\begin{aligned} t &\sim \text{Gamma}(a_t, b_t) \\ p(\sigma) &\sim 1/(\sigma(1-\sigma)) \\ p(c) &\sim 1/c \end{aligned} \tag{39}$$

This choice allows us to update t, σ and c through the Metropolis-Hastings proposals with log-Normal distributions.

We propose for $(x_i)_{i=1,\ldots,L}$ and γ the following prior distributions:

$$x_i \overset{iid}{\sim} \text{Uniform}[0, x_{\max}] \quad i = 1, \ldots, L \tag{40}$$

$$\gamma \sim \text{Gamma}(a_\gamma, b_\gamma).$$

We opt for another round of Metropolis-Hastings for the joint update of x and γ, using as proposal respectively a truncated Normal distribution on the same support and a log-Normal distribution.

We use as prior for the weights the exponentially tilted BFRY distribution described in eq. (37). In case of a simple indirect graph, we impute the latent variables which describe the underlying directed multigraph:

$$\bar{n}_{ij}|z, \tilde{w} \sim \begin{cases} \delta_0 & \text{if } z_{ij} = 0 \\ \text{tPoisson}(2p_{ij}\tilde{w}_i\tilde{w}_j) & \text{if } z_{ij} = 1 \text{ and } i < j \\ \text{tPoisson}(\tilde{w}_i^2) & \text{if } z_{ii} = 1 \end{cases} \tag{41}$$

where tPoisson(λ) is a zero-truncated Poisson whose density is $e^{-\lambda}\lambda^y/(y!(1-e^{-\lambda}))$ for $y \in \mathbb{N}_+$. To obtain a conjugate structure for the conditional distribution of the weights, we augment the model with the independent variables

$$u_i|\tilde{w}_i \sim \text{Truncated Poisson}(\zeta\tilde{w}_i),\ i = 1, \dots, L \tag{42}$$

and obtain the following joint distribution:

$$\begin{aligned} &p\left((\bar{n}_{ij})_{i,j=1}^L, (z_{ij})_{i,j=1}^L, (\tilde{w}_i)_{i=1}^L, (u_i)_{i=1}^L, (x_i)_{i=1}^L\right) \\ &= \left(\prod_{\substack{i,j=1 \\ i\neq j}}^L \frac{(p_{ij}\tilde{w}_i\tilde{w}_j)^{\bar{n}_{ij}} e^{-p_{ij}\tilde{w}_i\tilde{w}_j}}{\bar{n}_{ij}!}\right) \times \left(\prod_i \frac{\sigma\tilde{w}_i^{-1-\sigma} e^{-c\tilde{w}_i} \cancel{(1-e^{-\zeta\tilde{w}_i})}}{\Gamma(1-\sigma)((\zeta+c)^\sigma - c^\sigma)} \times \frac{(\zeta\tilde{w}_i)^{u_i} e^{-\zeta\tilde{w}_i}}{u_i!\cancel{(1-e^{-\zeta\tilde{w}_i})}} \times \frac{1}{x_{\max}}\right). \end{aligned} \tag{43}$$

From this we infer the conjugate full conditional distributions of $\tilde{w}_i$, which allows for a Gibbs sampler when no self edges are allowed. In fact,

$$\begin{aligned} p(\tilde{w}_i|rest, (\tilde{w}_j)_{j\neq i}) &\propto \tilde{w}_i^{-1-\sigma+u_i+\sum_{j\neq i}\bar{n}_{ij}} e^{-\tilde{w}_i\left(c+\zeta+\sum_{j\neq i}p_{ij}\tilde{w}_j\right)} \\ &= \text{Gamma}(-\sigma + u_i + \sum_{j\neq i}\bar{n}_{ij}, c + \zeta + \sum_{j\neq i} p_{ij}\tilde{w}_j) \end{aligned}$$

The Gibbs sampler, though, is known to have bad mixing in presence of strong correlation among the variables. Another possibility is to opt for a Hamiltonian Monte Carlo (HMC) step that updates all the weights of the nodes at the same time and is not limited by self edges. For details on the general HMC algorithm see Neal (2011). To perform it, we need to be able to compute the derivative of the posterior of $\tilde{w}$, that in our case is

$$\frac{\partial \log p(\log\tilde{w}|rest)}{\partial \log\tilde{w}_i} = m_i + u_i - \sigma - \tilde{w}_i(\zeta + c) - \tilde{w}_i\sum_j p_{ij}\tilde{w}_j \tag{44}$$

where $m_i = \sum_j(\bar{n}_{ij} + \bar{n}_{ji})$.

In algorithm 3 we present the steps of our posterior inference algorithm. For more details we refer to appendix 9.4.

Algorithm 3 Posterior Inference Algorithm

1: Update $\tilde{w}_1, \dots, \tilde{w}_L$ given the rest with a Gibbs sampling or HMC step.
2: Update the parameters t, σ, c given the rest with a MH step.
3: Update $\bar{n}_{ij}$ and u_i given the rest according to their conditional distributions of eq. (41) and eq. (42).
4: Update $x_1, \dots, x_L$ and γ given the rest with a MH step.

The Gibbs sampler, the HMC for $\tilde{w}$, the MH for γ and x all have quadratic complexity in the number of nodes when γ is positive due to the update of the variable p_{ij}. A possibility to decrease the complexity would be to look at the thinning procedures exploited in section 4 or look at further approximations. We discuss this as future work in section 8.

Another challenge of space models is the non identifiability of locations. In fact, a unique matrix $(p_{ij})_{ij}$ could be generated by an infinite number of configurations $(x_i)_i$, due to the invariance of the distance function to translations, rotations and reflections. A possible workaround is to use the Procrustean transformation to rescale and centre the posterior samples with respect to a reference configuration. For details of this method we refer to Hoff et al. (2002).

6 Experiments on simulated data

We study the performance of the MCMC algorithm on data generated from the approximate model: we simulate a graph according to eq. (6) with weights sampled from the exponentially-tilted BFRY distribution, the approximation to the GG proposed in section 5.1 with hyperparameters $\sigma = 0.2$, $c = 1.5$ and truncation level $L = 10^4$, and locations sampled uniformly at random in $[0, 5]$. The remaining parameters are set to $t = 80, \gamma = 1$. The resulting graph is sparse, has approximately 700 active nodes and 4000 edges.

In this experiment, we fix γ and run three MCMC chains to estimate σ, c, t, the sociabilities $(\tilde{w}_i)_{i=1,\dots,L}$, locations $(x_i)_{i=1,\dots,L}$, counts $(n_{ij})_{i,j=1,\dots,L}$ and auxiliary variables $(u_i)_{i=1,\dots,L}$. We initialise the first chain at the true values of the variables and hyperparameters, while we randomly pick values according to the prior distributions for the second and third chains. Each chain ran for 10^6 iterations and it took almost 48 hours to run the chains in parallel (using the Python library multiprocessing) on a standard laptop (central processor unit at 2.3 GHz, dual-core). We fixed the location of the highest degree node to its true value. This does not solve completely the identifiability issue for locations (described in section 5), as it does not guarantee a unique solution for the locations, but makes them identifiable up to reflection with respect to the highest degree node. To completely solve the challenge of identifiability, we could use the Procrustean transformation described in Hoff et al. (2002) or fix multiple locations.

In fig. 6 we show the traceplots of the MCMC samples for t, σ, c. The chains are able to recover the true values of the parameters.

In fig. 7, we plot the 95% posterior credible intervals for the sociability parameters of the highest and lowest degrees nodes. Over the whole set of nodes, 95% of the true values fall in the associated credible interval.

Figure 8 represents the posterior of the spatial parameters. We represent the traceplot of the location of a node with high degree. In chain 1 we see the phenomenon of identifiability up to reflection: the fixed node has location approximately 2.8 and the two modes visited by the chain are equidistant with respect to it. To have a clearer idea of the estimation of x, we report a scatter plot illustrating the relation between the true values of the location and their posterior

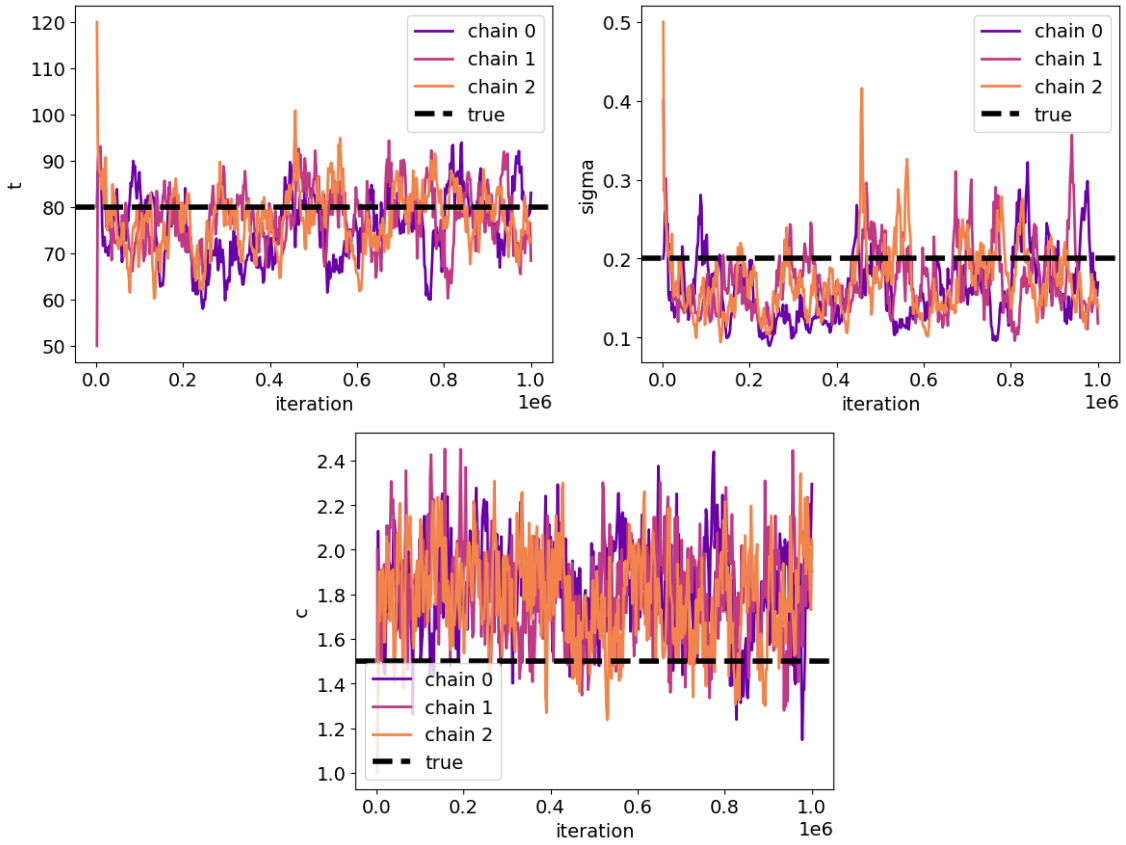


Figure 6: MCMC traceplot of t, σ and c for a graph simulated exponentially-tilted GBFRY sociabilities, uniform locations and $\sigma = 0.2, c = 1.5, t = 80, \gamma = 1, x_{\max} = 5$.

mean estimates. We see a clear correlation between the true and the estimated value for nodes with positive degree.

In fig. 9 we plot the logarithm of the posterior against the true value.

Figure 10 represents the 95% posterior credible intervals of the degree distributions, obtained by sampling from the posterior predictive of the model for one of the chains. The credible bands cover the empirical degree distribution of the original data.

These experiments show that the posterior inference scheme is able to recover well the sociability structure, the spatial structure and the parameters of the model. Future work requires to explore different ways to estimate the locations and achieve a faster convergence through better mixing as γ increases. As the parameter that tunes the spatial effects increase, the locations get more correlated and the posterior gets more peaked and multimodal, making the convergence difficult. In fact, the move of the chain from a local optimum to a region of the space with higher likelihood would require a passage through configurations with much smaller probability. A possible solution would be to explore parallel tempering algorithms (Swendsen and Wang (1986)) that relieve this problem by allowing each chain to draw insights from the different configurations explored by the others and use this information to explore the space differently.

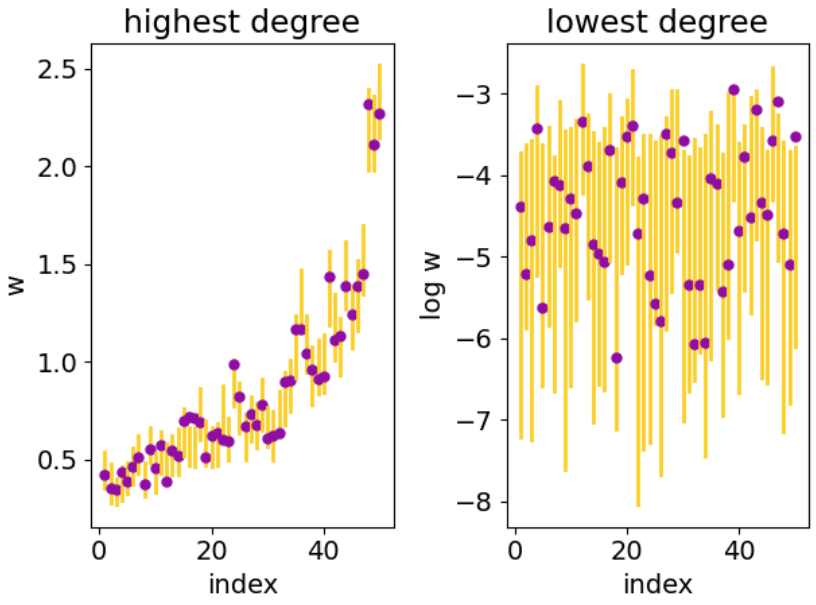


Figure 7: 95% posterior credible intervals of the 50 sociability parameters $\tilde{w}$ with the highest (on the left) and positive lowest values (on the right, in logarithmic scale) for a graph simulated with exponentially-tilted GBFRY sociabilities, uniform locations and $\sigma = 0.2, c = 1.5, t = 80, \gamma = 1, x_{\max} = 5$. The violet dot represents the true value., the yellow line is the credible interval. Overall, 95% of the true weights fall into their credible intervals.

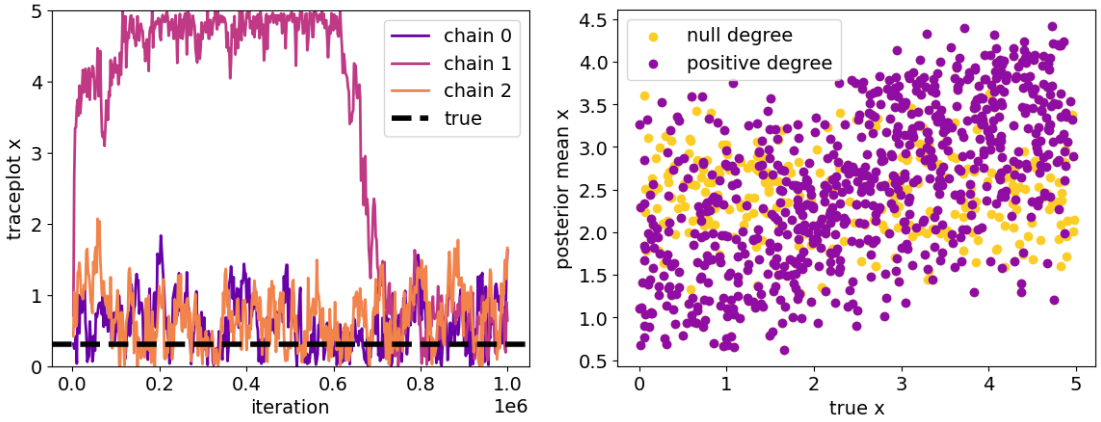


Figure 8: On the left MCMC traceplot of the location of a high degree node. On the right, scatter plot for one of the chains of the true location against the posterior mean of x (in yellow the nodes of null degree and in violet those with positive degree). The graph has been sampled with exponentially-tilted GBFRY sociabilities, uniform locations and $\sigma = 0.2, c = 1.5, t = 80, \gamma = 1, x_{\max} = 5$.

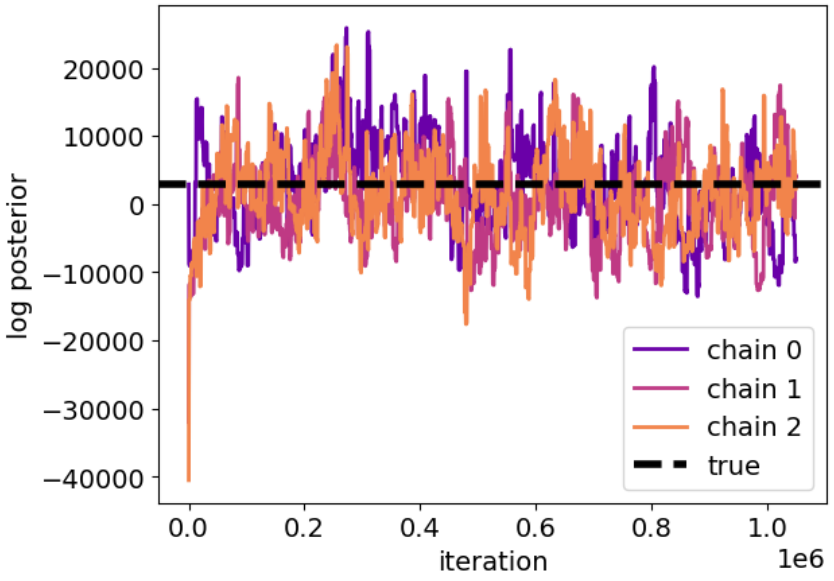


Figure 9: Traceplots of the log posteriors of the three chains. The graph has been sampled with exponentially-tilted GBFRY sociabilities, uniform locations and $\sigma = 0.2, c = 1.5, t = 80, \gamma = 1, x_{\max} = 5$.

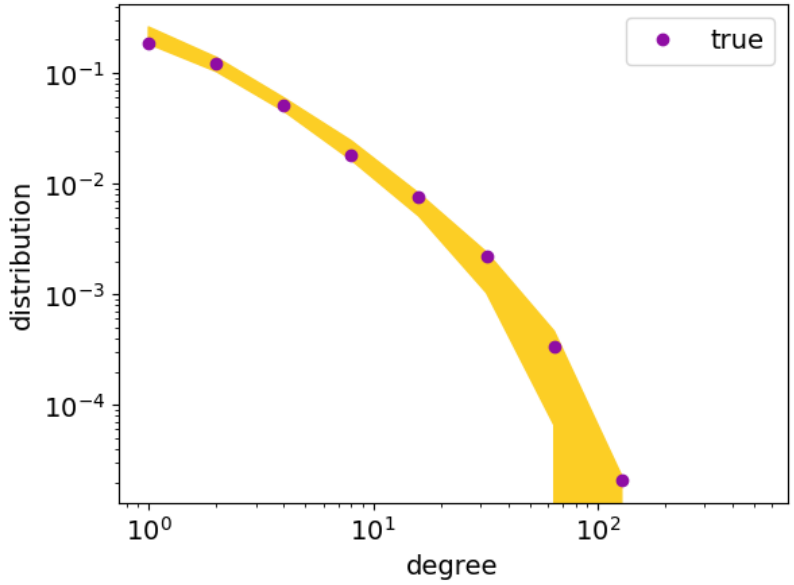


Figure 10: In yellow, 95% posterior credible intervals of the degree distribution. The dots represent the empirical distribution of the data, a graph sampled with exponentially-tilted GBFRY sociabilities, uniform locations and $\sigma = 0.2, c = 1.5, t = 80, \gamma = 1, x_{\max} = 5$.

7 Relations with other models

In this section we cover some of the spatial network models literature that can be related to our model. We reflect on similarities and differences and highlight how our proposal generalises some of them.

Homogeneous spatial random graph

Homogeneous random geometric graphs, also known as Gilbert-disk model as introduced by Gilbert (1961), are spatial models whose link probability depends only on the distance between nodes. Our proposal becomes equivalent to it when we set $W(\vartheta, \vartheta') = W'(|x - x'|)$.

Scale-free percolation (SFP) models

While we did not approach the study of percolation, many spatial network models focus on this property. The study of percolation regards the growth of the biggest connected component in the graph. In the limit, as the number of nodes goes to infinity, the probability of observing an infinite connected component is tuned by a parameter λ of which there exists a critical value λ_c such that for each $\lambda \geq \lambda_c$ we observe an infinite connected component with positive probability. Spatial scale-free percolation models have been around for at least 30 years (see Zhang et al. (1983)), but recently Deijfen et al. (2013) proposed a percolation model with locations drawn uniformly at random on the lattice $\mathbb{Z}^d$ and with heterogeneity of nodes given by random variables $w \in \mathbb{R}^d$ drawn from a distribution with regularly varying behaviour at infinity. The probability of connection is $1 - e^{-\lambda w_x w_y ||x-y||^{-\gamma}}$. For certain values of γ and of the regular variation exponent α it is shown that this model achieves a positive average local clustering, is small world or ultra-small world and has regularly varying tails. This model was further extended by Deprez and Wüthrich (2018) on the continuum space $\mathbb{R}^d$, by drawing locations according to a homogeneous Poisson point process. While this model produces graphs that have an almost surely infinite number of nodes and edges, a finite graph can be obtained by restricting the space of locations to a finite box. Among other important results on percolation and small world properties, they prove that if $\min\{\gamma, \gamma\alpha\} > d$, then the degree distribution has power-law tails with exponent $\tau := \gamma\alpha/d > 1$. For $\min\{\gamma, \gamma\alpha\} < d$, instead, the degree distribution is degenerate having almost surely infinite expected value. For the same model, Dalmau and Salvi (2019) proved that for $\gamma > d$ and $\tau > 1$ the average local clustering coefficient exists, does not depend on the locations' process and is strictly positive.

Considering the vast parallelism between the scale-free in continuum space and our model, we might wonder if it possible to extend these results to our case. In fact, we provide asymptotic results in section 3 only for the time t that goes to infinity, but we do not study the asymptotic behavior in space ($x_{\max} \to \infty$). The main difference between the two models is that the weights in the scale-free percolation model are drawn from a distribution, while in our case they come from a CRM. In the case of infinite activity CRM (the most interesting to us), even if we restrict the covariates x and the labels θ on finite windows the number of potential nodes is still infinite. In the scale-free percolation model, instead, it suffices to restrict the domain of the point process on a finite box and automatically the number of nodes (with null or positive degree) will be almost surely finite. This is a substantial difference and does not allow for a trivial extension of their results to our model. We conjecture that some of the results might hold in our case as well, and we leave this to future work.

Hyperbolic random graph and GIRG

The extended geometric heterogeneous random graph (extended GIRG) model proposed by Komjáthy and Lodewijks (2019) is obtained by drawing locations x from a homogeneous Poisson process on $\mathbb{R}_+$ with intensity $\nu > 0$, and weights w from a distribution P and connecting two nodes i and j with probability $h(w_i, w_j, |x_i - x_j|)$, $h : \mathbb{R}_+^3 \to [0,1]$ to be further specified. We can report this model to our framework by setting as graphon function $W(\vartheta, \vartheta') := h(g(u), g(u'), |x - x'|)\mathbb{1}_{u<1}\mathbb{1}_{u'<1}$, where g is the generalised inverse of $1 - P$. Given that g has bounded support, W inherits this property and therefore the extended GIRG model generates only dense or empty graphs.

The extended GIRG, as the name suggests, was born as an extension of the GIRG proposed by Bringmann et al. (2019), which is itself a reparametrisation of the hyperbolic random graph (HRG) by Krioukov et al. (2010). The HRG is a homogeneous spatial random graph model whose locations are sampled uniformly at random in the hyperbolic space $\mathbb{H}^d$ and, in the simplest version, an edge is drawn whenever the distance between two nodes is less than a threshold. While formally introduced in Krioukov et al. (2010), it was empirically shown to be useful in Boguñá et al. (2010) when the hyperbolic space was used as the embedding space of the internet graph. While many other embeddings have been proposed, especially in Euclidean spaces, a simple hyperbolic mapping was shown to naturally retain many of the important characteristics of the original network, in particular its scale-free degree distribution, the average degree and the clustering coefficient. The GIRG by Bringmann et al. (2019) contains the HRG as a particular case, by mapping the hyperbolic coordinates of nodes into weights in $\mathbb{R}_+$ and locations in $[-n/2, n/2]$ (n number of nodes). Later, Komjáthy and Lodewijks (2019) generated the extended GIRG by proving that from in the asymptotics regime of GIRG weights are distributed as power-laws and locations are drawn from a homogeneous Poisson process.

Note also that the GIRG and the SFP are very similar models, which differ mainly in the parametrization and the scope for which they were initially studied. Originally, they differed in the space of locations, that for Deijfen et al. (2013) was $\mathbb{Z}^d$ while in GIRG has always been $\mathbb{R}^d$, and in the probability of connection. These differences, though, disappeared thanks to the extensions offered by Deprez and Wüthrich (2018) and Komjáthy and Lodewijks (2019).

Sparse latent space model

Spencer and Shalizi (2017) present the rectangular latent positions model (rectangular LPM) to achieve a sparse, projective and learnable spatial networks, which belongs to the more general class of latent position models by Hoff et al. (2002). In this new framework, points (x, r) are drawn from a homogeneous Poisson process on $\mathbb{R}^d \times \mathbb{R}$. Given two points $x, y \in \mathbb{R}^d$ the probability of connection depends solely on the locations through the function $K(\|x - y\|)$, where $K : \mathbb{R}^d \to [0,1]$ is a link probability function. To get a finite graph it is once again necessary to restrict the space through a function $H : \mathbb{R}_+ \to \mathcal{B}(\mathbb{R}^d) \times \mathcal{B}(\mathbb{R})$, with $\mathcal{B}(\mathbb{R})$ and $\mathcal{B}(\mathbb{R}^d)$ the Borel $\sigma-$algebras, such that for $t_1 < t_2$ we have $H(t_1) \subset H(t_2)$ and $|H(t)| = t$. They focus on the special case of the rectangular LPM by setting $H(t) = [-g(t), g(t)]^d \times \left[0, t/(2g(t))^d\right]$, where $g(t) := t^{p/d}$, $p \in [0,1]$. In this case, they prove that $N_t^{(e)} \sim N_t^{2-p}$. Without this auxiliary space construction, a linear increase of $H(t)$ would imply $N_t^{(e)} \sim N_t$, lacking flexibility in sparsity. The resulting model is projective and learnable (i.e. they provide upper bounds for speed of convergence of the estimators of the latent positions, squared latent distance and link probability).

They provide comparisons with the sparse graphon and the graphex frameworks. The sparse

graphon (Bollobás and Riordan (2009), Borgs et al. (2019)) is not projective, but they prove it to be learnable. A projective model is instead the classic graphex (Caron and Fox (2017), Veitch and Roy (2015)), whose learnability is a difficult open problem. To compare the two models, consider the basic graphex process (without covariates) and set ϑs to be the locations and θs to be the auxiliary variables. The link probability function is more general in the graphex framework than it is in the rectangular LPM, even if it could be extended in the latter at extra cost. A naive transposition of the graphex to the rectangular LPM is still not possible since, if W was uniquely dependent on locations, we could never satisfy the integrability condition of W, necessary to observe a finite number of edges. With our spatial proposal, we enrich with additional node variables the original proposal of Caron and Fox (2017), obtaining a more direct way to embed spatial information. In fact, while in the base model the weights $w = \bar{\rho}^{-1}(\vartheta)$ could be interpreted as latent locations, their interpretation as sociability parameters allows the use of regular variation (which would not be very natural with spatial locations), offering all the desirable sparsity and scale-free properties. These, in fact, are easily tuned by α, the exponent of the regular variation of the Lévy measure.

8 Conclusion

In this work we present a novel model for sparse spatial random graphs. The model is inhomogeneous as each node is associated with two variables: a sociability weight and a more general variable which can be conveniently interpreted as a location (real or latent). We select a link probability function that is increasing in the sociabilities and decreasing in the distance between two nodes. The model fits into the more general graphex framework by generalising with a spatial component the proposal of Caron and Fox (2017). This, together with the Bayesian nonparametric construction based on completely random measures, allows our model to inherit flexibility and interpretability of the parameters. In particular, through the distribution of the sociabilities we are able to tune the sparsity of the connections and achieve different power-law degree distributions. We also prove asymptotic results on the number of nodes, edges and clustering coefficient. Interestingly, we show that the clustering coefficient is bounded away from zero even in the non spatial version of the model. All our asymptotic results are in the time domain, while further work would be required to assess the asymptotics in space and compare the theorems with the broader space networks literature. Future work in this direction would also include a discussion about the small or large world properties of the model.

While our study focused on the use of the generalised gamma process to achieve a power-law for small degree nodes, another proposal could be advanced. Li and Cai (2004) and Paleari et al. (2010) observed a double power-law degree distribution behavior in the networks of airports connections around the world and Seshadri et al. (2008) inferred the same behaviour for mobile calls. This means that nodes with low and high degrees exhibit power-law behaviours with different exponents. To achieve it, we could consider a marked Poisson process with Lévy measure doubly regularly varying, both at zero and infinity, for example the generalised gamma Pareto process proposed in Ayed et al. (2019).

In terms of computational aspects, we propose different simulation algorithms for finite activity or truncated infinity activity CRMs, reducing the computational complexity from the naive $O(N^2)$ to $O(N\log^2(N))$. We also explain how to simulate exactly without truncations an infinite activity CRM. We propose a posterior inference algorithm that targets an approximation of the posterior. It relies on a combination of Gibbs, Hamiltonian Monte Carlo and Metropolis

Hastings steps. We complement this algorithms with studies on simulated data to show the convergence to the true posterior, and the next step requires to test the model with real spatial network data.

The computational complexity of the algorithm is quadratic in the number of nodes and demands further work to devise a smart way to scale it down. A possible workaround would be to explore further approximations of the likelihood. A proposal in this direction has been advanced in Rastelli et al. (2018) to solve a similar problem in the latent space model of Hoff et al. (2002): after defining a grid on the space $B_1, \dots, B_K$ (as in section 4) they approximate the likelihood by simply computing it at the centre of each box. Their suggestion cannot be directly applied to our framework because of the term $\sum_{i,j} w_i w_j / (1 + |x_i - x_j|)^\gamma$ in eq. (43), which mixes w and x. We would therefore need to integrate it with the weight layers used in section 4. This would add other layers of approximation to our algorithm and would require further investigation to understand the goodness of it. Another possibility would be to exploit the technique of Poisson-minibatching, which requires to augment the model with auxiliary variables whose presence reduces the complexity required to compute the augmented joint likelihood. This approach does not require approximations of the likelihood and leads to exact inference. Details of a more general procedure can be found in Zhang and De Sa (2019).

Eventually, we discuss our contribution in the vast literature of spatial random graph modes, drawing a map of models that are connected to ours and discussing how the sparse spatial random graph model is linked to them and sometimes able to generalise them.

9 Appendix

9.1 Proofs

Proof of theorem 3.2. The proofs follows noting that the probability of a connection is bounded from above and below:

$$1 - e^{-\frac{2w_i w_j}{(1+x_{\max})^\gamma}} \leq 1 - e^{-\frac{2w_i w_j}{(1+|x_i - x_j|)^\gamma}} \leq 1 - e^{-2w_i w_j}$$

The upper bound corresponds exactly to Caron and Fox (2017) model and the lower bound is a simple reparametrization of it with $\tilde{w} := \frac{w}{\sqrt{(1+x_{\max})^\gamma}}$ which does not affect the activity of the process. The density level of the two extreme processes depend solely on the activity of the underlying CRM, hence the graphs resulting from the lower and upper bound are going to be either both sparse or both dense. Therefore, defining $\tilde{N}_t$, $\tilde{N}_t^{(e)}$ and $\bar{N}_t$, $\bar{N}_t^{(e)}$ the number of nodes and edges of respectively the lower and upper bound process, we have that

$$\Theta(t^2) = \Theta(\tilde{N}_t^2) = \tilde{N}_t^{(e)} \leq N_t^{(e)} \leq \bar{N}_t^{(e)} = \Theta(\bar{N}_t^2) = \Theta(t^2)$$

where the first and last equalities come from Theorem 2 of Caron and Fox (2017). Therefore, $N_t^{(e)} = \Theta(t^2)$. The same reasoning applies to N_t, and following Caron and Fox (2017) we get that in the finite activity case $N_t = \Theta(t)$ and in the infinite activity case $N_t = O(t)$, which proves the result. □

Proof of theorem 3.4. Model of eq. (6) can be rewritten as

$$W((u_i, x_i), (u_j, x_j)) = 1 - e^{-\eta(u_i, u_j)\omega(x_i, x_j)}, \tag{45}$$

where $\eta(u_i, u_j) = 2\bar{\rho}^{-1}(u_i)\bar{\rho}^{-1}(u_j)$ and $\omega(x_i, x_j) = \frac{1}{(1+|x_i - x_j|)^\gamma}$. We are now in the framework of Caron et al. (2020) and can apply their Remark 26, obtaining

$$N_t^{(e)} \sim \begin{cases} N_t^{\frac{2}{1+\alpha}} & \text{for } \alpha \in (0,1) \\ \frac{N_t^2}{\log(N_t)^2} & \text{for } \alpha = 0 \\ N_t^2 & \text{for } \alpha = 1 \end{cases} \tag{46}$$

Moreover, consider the function $\mu(u, x) := \int_{\mathbb{R}_+} \int_F W((u, x), (v, y)) dv dy$. It holds that

$$C_1 \mu(u)\mu(x) \leq \mu(u, x) \leq C_2 \mu(u)\mu(x)$$

with $\mu(u) = \int_{\mathbb{R}_+} \eta(u, u_j) du_j$ and $\mu(x) = \int_F \omega(x, x_j) dx_j$, for some positive constants C_1, C_2. We can then apply Proposition 17 in Caron et al. (2020) to the process described by the graphon function $W((u_i, x_i), (u_j, x_j)) = \eta(u, u_j)\omega(x, x_j)$ and obtain an almost sure convergence of the proportion of nodes with degree j for it. We can then bound the same quantity for our original process obtaining for $\alpha \in (0, 1)$:

$$\frac{C_1}{C_2} \frac{\sigma\Gamma(j - \alpha)}{j!\Gamma(1 - \alpha)} \leq \lim_{t \to \infty} \frac{N_{t,j}}{N_t} \leq \frac{C_2}{C_1} \frac{\alpha\Gamma(j - \alpha)}{j!\Gamma(1 - \alpha)}, \text{ for } j \geq 1.$$

For $\alpha \leq 0$ we get

$$\frac{N_{t,j}}{N_t} \to 0, \text{ for } j \geq 1 \text{ a.s..}$$

□

Proof of theorem 3.5. Call $T_t := \frac{1}{6}\sum_{i\neq j\neq k} z_{ij}z_{ik}z_{jk}\mathbb{1}_{\theta_i\leq t}\mathbb{1}_{\theta_j\leq t}\mathbb{1}_{\theta_k\leq t}$ the number of triangles and $A_t := \frac{1}{2}\sum_{i\neq j\neq k} z_{ij}z_{ik}\mathbb{1}_{\theta_i\leq t}\mathbb{1}_{\theta_j\leq t}\mathbb{1}_{\theta_k\leq t}$ the number of open and closed triplets in the graph. By definition 21, the global clustering coefficient can be written as $C_t^{(g)} = 3T_t/A_t$ that is, the number of triangles over the number of open and closed triples. We now exploit Proposition 56 of Borgs et al. (2019), since our model can be seen as a special case of theirs. From that, we know that

$$3T_t \sim t^3 \int W((w_i,x_i),(w_j,x_j))W((w_i,x_i),(w_k,x_k))W((w_k,x_k),(w_j,x_j))\rho(dw_i)\rho(dw_j)\rho(dw_k)d\mathbf{x}$$

and

$$A_t \sim t^3 \int W((w_i,x_i),(w_j,x_j))W((w_i,x_i),(w_k,x_k))\rho(dw_i)\rho(dw_j)\rho(dw_k)d\mathbf{x}$$

from which we get our result by dividing the two quantities.

By adapting Proposition 10 in Caron et al. (2020) to our graphon function of eq. (6) we get that the local clustering coefficient converges in probability to

$$C_t^{(l)} \to \lim_{w\to 0} \frac{\int(1-e^{-\frac{2ww_j}{(1+|x_i-x_j|)^\gamma}})(1-e^{-\frac{2ww_k}{(1+|x_i-x_k|)^\gamma}})(1-e^{-\frac{2w_jw_k}{(1+|x_j-x_k|)^\gamma}})\rho(dw_j)\rho(dw_k)dx_idx_jdx_k}{(\int(1-e^{-\frac{2ww_j}{(1+|x_i-x_j|)^\gamma}})\rho(dw_j)dx_idx_j)^2}$$

Using the inequality $1-e^{-tw} \leq tw$ we can employ dominated convergence theorem and exchange limits and integrals. Exploiting the equivalence $1-e^{-tw} \sim tw$ as $t\to 0$ and substituting $w = \bar{\rho}^{-1}(\vartheta)$, the numerator behaves as

$$\begin{aligned}&\int(1-e^{-\frac{2\bar{\rho}^{-1}(\vartheta)w_j}{(1+|x_i-x_j|)^\gamma}})(1-e^{-\frac{2\bar{\rho}^{-1}(\vartheta)w_k}{(1+|x_i-x_k|)^\gamma}})(1-e^{-\frac{2w_jw_k}{(1+|x_j-x_k|)^\gamma}})\rho(dw_j)\rho(dw_k)dx_idx_j\\ \sim& 4\bar{\rho}^{-1}(\vartheta)^2\int\frac{w_jw_k}{(1+|x_i-x_j|)^\gamma(1+|x_i-x_k|)^\gamma}(1-e^{-\frac{2w_jw_k}{(1+|x_j-x_k|)^\gamma}})\rho(dw_j)\rho(dw_k)dx_idx_jdx_k\end{aligned}$$

For the denominator, we use again dominated convergence and the same equivalence to get

$$\left(\int(1-e^{-\frac{2\bar{\rho}^{-1}(\vartheta)w_j}{(1+|x_i-x_j|)^\gamma}})\rho(dw_j)dx_idx_jdx_k\right)^2 \sim 4\bar{\rho}^{-1}(\vartheta)^2\left(\int\frac{w_j}{(1+|x_i-x_j|)^\gamma}\rho(dw_j)dx_idx_j\right)^2$$

Dividing numerator and denominator we get the result. □

9.2 Details and complexity of algorithm 2

Details of algorithm 2

Consider the partition of nodes into cells acccording to their locations, proposed in section 4.1. Note that we can split eq. (31) over the constructed grid in the following way:

$$\mu_{t,x_{\max}} = \underbrace{\sum_{k=1}^{K}\sum_{l=k,k+1}\sum_{i:x_i\in A_k}\sum_{j\geq i:,x_j\in A_l}\frac{2w_iw_j}{(1+|x_i-x_j|)^\gamma}\delta_{(\theta_i,\theta_j,x_i,x_j)}}_{=:\mu^{(1)}_{t,x_{\max}}}$$

$$+\underbrace{\sum_{k=1}^{K}\sum_{\ell\geq k+2}\sum_{l=k,k+1}\sum_{i:x_i\in A_k}\sum_{j\geq i:,x_j\in A_l}\frac{2w_iw_j}{(1+|x_i-x_j|)^\gamma}\delta_{(\theta_i,\theta_j,x_i,x_j)}}_{=:\mu^{(2)}_{t,x_{\max}}}$$

The connections coming from $\mu^{(1)}_{t,x_{\max}}$ can be simulated with a time complexity linear in N using the standard simulation technique explained at the beginning of section 4. In fact, the number of nodes in every cell is $O(1)$ (as K grows at most as N) and we have $O(N)$ cells and the computational time is therefore $O(N)$.

This space partition, though, is not enough to have a big gain in the computational complexity, hence we add the weights partition into layers proposed in section 4.1. We can partition the measure $\mu^{(2)}_{t,x_{\max}}$ on every pair of weight layers V_{j_1}, V_{j_2} and every pair of cells A_k, A_l obtaining a restricted measure $\mu^{(j_1j_2)}_{t,x_{\max}}$. We rewrite and upper bound $\mu^{(j_1j_2)}_{t,x_{\max}}$ to rely on thinning properties of Poisson processes.

$$\begin{aligned}
\mu^{(j_1j_2)}_{t,x_{\max}} &= \sum_{k=1}^{K-2}\sum_{l=k+2}^{K}\frac{\overline{w}^{(j_1)}_k\overline{w}^{(j_2)}_l}{(1+\delta(l-k)-\delta)^\gamma}\sum_{i\in A_k\cap V_{j_1}}\sum_{j\in A_l\cap V_{j_2}}\frac{w_i}{\overline{w}^{(j_1)}_k}\frac{w_j}{\overline{w}^{(j_2)}_l}\frac{(1+\delta(l-k)-\delta)^\gamma}{(1+|x_i-x_j|)^\gamma}\delta_{(\theta_i,\theta_j,x_i,x_j)}\\
&\leq \sum_{k=1}^{K-2}\overline{w}^{(j_1)}_k\underline{w}_{j_2+1}|V_{j_2}|C_{k,K,\gamma,\delta}\sum_{l=k+2}^{K}p(l-k;k,K,\gamma,\delta)\\
&\quad\sum_{\substack{i\in A_k\cap V_{j_1}\\ j\in A_l\cap V_{j_2}}}\frac{w_i}{\overline{w}^{(j_1)}_k}\frac{w_j}{\overline{w}^{(j_2)}_l}\frac{(1+\delta(l-k)-\delta)^\gamma}{(1+|x_i-x_j|)^\gamma}\delta_{(\theta_i,\theta_j,x_i,x_j)}
\end{aligned}$$

where $p(l;k,K,\gamma,\delta) = 1/(C_{k,K,\gamma,\delta}(1+(-\delta+\delta l))^\gamma)$ is the probability mass function of a Zipf distribution on $\{k+2,\dots,K\}$ of which $C_{k,K,\gamma,\delta} := \sum_{l=k+2}^{K}(1-\delta+\delta|k-l|)^{-\gamma}$ is the normalizing constant. For simplicity, let us call $p(\cdot) := p(\cdot;k,K,\gamma,\delta)$.

We can now use thinning properties of Poisson processes to obtain the sampling procedure from $\mu^{(2)}_{t,x_{\max}}$ illustrated in algorithm 2.

Complexity of algorithm 2

We start computing the expected value of the number of edges having an endpoint in cell k, sum over all cells k and then over all weight layers j_1, j_2. Summing over k and calling, for simplicity, $\tilde{C} := \sum_{k=1}^{K}C_{k,K,\gamma,\delta}$:

$$\begin{aligned}
\mathbb{E}\left[\sum_{k=1}^{K}m^{(j_1j_2)}_k\right] &= \mathbb{E}\left[\sum_{k=1}^{K}\overline{w}^{(j_1)}_k\underline{w}_{j_2+1}|V_{j_2}|C_{k,K,\gamma,\delta}\right] \leq \tilde{C}\underline{w}_{j_2+1}\mathbb{E}\left[|V_{j_2}|\sum_i w_i\mathbb{1}_{i\in w_{j_1}}\right]\\
&= \tilde{C}\underline{w}_{j_2+1}\int_{\underline{w}_{j_2}}^{\underline{w}_{j_2+1}}\rho(dw)\int_{\underline{w}_{j_1}}^{\underline{w}_{j_1+1}}w\rho(dw)\\
&= \begin{cases}
\tilde{C}\underline{w}_{j_2+1}^{1-\sigma}\underline{w}_{j_1+1}^{1-\sigma} & \text{if } \underline{w}_{j_2+1}\leq 1 \text{ and } \underline{w}_{j_1+1}\leq 1\\
\tilde{C}\underline{w}_{j_2+1}^{1-\sigma}\underline{w}_{j_1+1}^{1-\tau} & \text{if } \underline{w}_{j_2+1}\leq 1 \text{ and } \underline{w}_{j_1} > 1\\
\tilde{C}\underline{w}_{j_2+1}^{1-\tau}\underline{w}_{j_1+1}^{1-\sigma} & \text{if } \underline{w}_{j_2} > 1 \text{ and } \underline{w}_{j_1+1}\leq 1\\
\tilde{C}\underline{w}_{j_2+1}^{1-\tau}\underline{w}_{j_1+1}^{1-\tau} & \text{if } \underline{w}_{j_2} > 1 \text{ and } \underline{w}_{j_1} > 1
\end{cases}
\end{aligned}$$

We are omitting the cases in between (e.g. $\underline{w}_{j_1} \leq 1, \underline{w}_{j_1+1} > 1$), but the results hold the same. Let J be the index such that $\underline{w}_{J+1} \geq 1$ and $\underline{w}_J < 1$. Let us now define $m^{(j_1,j_2)} := \sum_{k=1}^{K} m_k^{(j_1,j_2)}$ and sum over all possible j_1, j_2 in expectation:

$$\begin{aligned}
\mathbb{E}\left[\sum_{j_1=1}^{\bar{J}}\sum_{j_2=1}^{\bar{J}} m^{(j_1,j_2)}\right] \leq & \tilde{C}\sum_{j_1=1}^{J}\underline{w}_{j_1+1}^{1-\sigma}\sum_{j_2=1}^{J}\underline{w}_{j_2+1}^{1-\sigma} + \tilde{C}\sum_{j_1=1}^{J}\underline{w}_{j_1+1}^{1-\sigma}\sum_{j_2=J+1}^{\bar{J}}\underline{w}_{j_2+1}^{1-\tau} + \\
& + \tilde{C}\sum_{j_1=J+1}^{\bar{J}}\underline{w}_{j_1+1}^{1-\tau}\sum_{j_2=1}^{J}\underline{w}_{j_2+1}^{1-\sigma} + \tilde{C}\sum_{j_1=J+1}^{\bar{J}}\underline{w}_{j_1+1}^{1-\tau}\sum_{j_2=J+1}^{\bar{J}}\underline{w}_{j_2+1}^{1-\tau} \\
= & \left(\tilde{C}\sum_{j=1}^{J}\underline{w}_{j+1}^{1-\sigma} + \tilde{C}\sum_{j=J+1}^{\bar{J}}\underline{w}_{j+1}^{1-\tau}\right)^2 \\
= & \tilde{C}^2\left(\underline{w}_0^{1-\sigma}2^{J(1-\sigma)} + \underline{w}_0^{1-\tau}2^{(\bar{J}-J)(1-\tau)}\right)^2 \\
< & \tilde{C}^2\left(1^{1-\sigma} + \underline{w}_0^{1-\tau}2^{(\bar{J}-J)(1-\tau)}\right)^2. \qquad (47)
\end{aligned}$$

Note that $C_{k,K,\gamma,x_{\max}} = \sum_{l=k+2}^{K}\frac{1}{(1-\delta+\delta|k-\ell|)^\gamma} \leq \sum_{l\geq 0}\frac{1}{(1+\delta(l-2)-\delta)^\gamma} \leq \frac{1}{\delta^\gamma}\sum_{l\geq 0}\frac{1}{(l-2)^\gamma}$ and this last series converges as long as $\gamma > 1$ (p-series). Hence, $\tilde{C} = O(1)$. Moreover, $\tilde{C}$ can be computed in linear time in K as the following recursion holds: $C_{k,K,\gamma,\delta} = C_{k-1,K,\gamma,\delta} + \frac{1}{(1-\delta+\delta(K-k))^\gamma}$. Since $\tilde{C} = O(1)$, this bound is $O(1)$ as long as $\tau > 1$.

We are finally able to compute the complexity of algorithm 2. Lines 1 to 7 can be executed in linear time with respect to N. In particular for line 5, since $\bar{J} = O(\log N)$, with counting sort or bucket sort we can determine the weight layers in time $O(N)$. The for loops in lines 8 to 12 take time $\sum_{j_1}\sum_{l}\Theta(|V_{j_1} \cap A_l|) = \Theta(N)$, as $|V_{j_1} \cap A_l|\, j_1 = 1, \ldots, \bar{J}, l = 1, \ldots, K$ form a partition of the space of all nodes. As for the for loops in lines 13 to 28, the complexity is $\sum_{j_1}\sum_{j_2}\sum_{k} O(1) = O(N\log^2 N)$ since we showed in eq. (47) that on average $\sum_{j_1}\sum_{j_2}\sum_{k} m_k^{(j_1,j_2)} = O(1)$. Hence, the overall complexity is $O(N\log^2 N)$.

9.3 Approximations of CRMs

Consider the Poisson random measure N on the space $(0, +\infty) \times \Theta$ with intensity measure $\nu(dw, d\theta) = \rho(dw)\xi_w(d\theta)$, with μ_w a Markov probability kernel from $(0, \infty)$ to Θ and ρ a Borel measure such that

$$\int_0^\infty (1 - e^{-w})\rho(dw) < \infty \text{ and } \int_0^\infty \rho(dw) = \infty.$$

Note that in our setting ξ_w does not depend on w. The functional

$$A = \int_0^\infty wN(dw, d\theta) \qquad (48)$$

is an infinite activity completely random measure on Θ with random weights and atoms. A can also be expressed in the form

$$A = \sum_{i\geq 1} w_i\delta_{\theta_i} \qquad (49)$$

where $(w_i)_{i\geq 1}$ are stochastically ordered, i.e. for every $w > 0$ we have that $\mathbb{P}(w_{i+1} > w) \leq \mathbb{P}(w_i > w)$. Let A_L be the CRM truncated after L steps:

$$A_L = \sum_{i=1}^{L} w_i \delta_{\theta_i} = \sum_{i=1}^{L} \bar{w}_{L,i} \delta_{\bar{\theta}_{L,i}}. \tag{50}$$

where $(\bar{w}_{L,i})_{i=1\dots L}$ is a finitely exchangeable random sequence such that $\bar{w}_{L,i} = w_{\pi_n(i)}$, with π_n permutation of the labels in the set $[n]$. Following Lee (2019), a finite independent and identically distributed approximation of the truncated CRM of eq. (50) is

$$\tilde{A}_L = \sum_{i=1}^{L} \tilde{w}_{L,i} \delta_{\tilde{\theta}_{L,i}} \tag{51}$$

where

$$\begin{aligned} \tilde{w}_{L,i} &\overset{iid}{\sim} \tilde{f}_L = f_{\Psi^{-1}(L)}(dw) = \frac{\Lambda_w(\Psi^{-1}(L))\rho(dw)}{L} \\ \tilde{\theta}_{L,i} | \tilde{w}_{L,i} = \tilde{w}_{L,i} &\sim \xi_{\tilde{w}_{L,i}} \end{aligned} \tag{52}$$

where $\Lambda_w(t) = \int_0^t \lambda_w(du)$ is the distribution function of a Markov probability kernel $\lambda_w(dt)$ from $(0,\infty)$ to $(0,\infty)$, and $\Psi(t) = \int_0^\infty \Lambda_w(t)\rho(dw)$.

Theorem 9.1 (Lee (2019))**.** *Let $\tilde{A}_L$ be the finite independent and identically distributed approximation defined by eq. (51) and eq. (52). Then, $\tilde{A}_L$ converges in distribution to the homogeneous CRM A with intensity $\rho(dw)\xi_w(d\theta)$.*

9.4 Details on the MCMC algorithm for the approximate generalised gamma process

Metropolis-Hastings for t, σ, c

The posteriors of t, σ and c are intractable, hence the first part of our posterior inference algorithm relies on Metropolis Hastings proposals for t, σ and c, following Caron and Fox (2017). In particular, we will use as proposals:

$$\begin{aligned} \frac{\tilde{\sigma}}{1-\tilde{\sigma}} | \sigma &\sim \text{log Normal}\left(\log \frac{\sigma}{1-\sigma}, \sigma_\sigma^2\right) \\ \tilde{c} | c &\sim \text{log Normal}(\log(c), \sigma_c^2) \\ \tilde{t} | t &\sim \text{log Normal}(\log(t), \sigma_t^2) \end{aligned}$$

where log Normal(μ, σ^2) is a log Normal distribution with mean μ and standard deviation σ. The change of variables $\frac{\tilde{\sigma}}{1-\tilde{\sigma}} = e^X$, where $X \sim \text{Normal}\left(\log \frac{\sigma}{1-\sigma}, \sigma_\sigma^2\right)$ implies the following density function for $\tilde{\sigma}$:

$$q_{\tilde{\sigma}}(\tilde{\sigma}) = p_X\left(\log\left(\frac{\tilde{\sigma}}{1-\tilde{\sigma}}\right)\right) \left| d\left(\log\left(\frac{\tilde{\sigma}}{1-\tilde{\sigma}}\right)\right)/d\tilde{\sigma} \right| \tag{53}$$

$$= \frac{1}{\sqrt{2\pi\sigma_\sigma^2}} e^{-\frac{(\log(\tilde{\sigma}/(1-\tilde{\sigma})) - \log(\sigma/(1-\sigma)))^2}{2\sigma_\sigma^2}} \frac{1}{\tilde{\sigma}(1-\tilde{\sigma})}. \tag{54}$$

Combining the complete likelihood, the priors and the proposals we get an acceptance ratio that is the minimum between 1 and r, r being

$$
\begin{aligned}
r = & \frac{\left(\prod_{i=1}^{L} \frac{\tilde{\sigma}\tilde{w}_i^{-1-\tilde{\sigma}} e^{-\tilde{c}\tilde{w}_i}(1-e^{-\tilde{\zeta}\tilde{w}_i})}{\Gamma(1-\tilde{\sigma})((\tilde{\zeta}+\tilde{c})^{\tilde{\sigma}}-\tilde{c}^{\tilde{\sigma}})} \times \frac{(\tilde{\zeta}\tilde{w}_i)^{u_i} e^{-\tilde{\zeta}\tilde{w}_i}}{u_i!(1-e^{-\tilde{\zeta}\tilde{w}_i})}\right) \frac{1}{\tilde{\sigma}(1-\tilde{\sigma})} \frac{1}{\tilde{c}} \frac{(b_t)^{a_t}}{\Gamma(a_t)} \tilde{t}^{a_t-1} e^{-b_t\tilde{t}}}{\left(\prod_{i=1}^{L} \frac{\sigma\tilde{w}_i^{-1-\sigma} e^{-c\tilde{w}_i}(1-e^{-\zeta\tilde{w}_i})}{\Gamma(1-\sigma)((\zeta+c)^{\sigma}-c^{\sigma})} \times \frac{(\zeta\tilde{w}_i)^{u_i} e^{-\zeta\tilde{w}_i}}{u_i!(1-e^{-\zeta\tilde{w}_i})}\right) \frac{1}{\sigma(1-\sigma)} \frac{1}{c} \frac{(b_t)^{a_t}}{\Gamma(a_t)} t^{a_t-1} e^{-b_t t}} \\
& \times \frac{e^{-\frac{(\log(\sigma/(1-\sigma))-\log(\tilde{\sigma}/(1-\tilde{\sigma})))^2}{2\sigma_\sigma^2}} \frac{e^{-\frac{(\log(c)-\log(\tilde{c}))^2}{2\sigma_c^2}}}{c\sigma_c} \frac{e^{-\frac{(\log(t)-\log(\tilde{t}))^2}{2\sigma_t^2}}}{t\sigma_t}}{e^{-\frac{(\log(\tilde{\sigma}/(1-\tilde{\sigma}))-\log(\sigma/(1-\sigma)))^2}{2\sigma_\sigma^2}} \frac{e^{-\frac{(\log(\tilde{c})-\log(c))^2}{2\sigma_c^2}}}{\tilde{c}\sigma_c} \frac{e^{-\frac{(\log(\tilde{t})-\log(t))^2}{2\sigma_t^2}}}{\tilde{t}\sigma_t}} \\
= & \frac{\tilde{\sigma}^L \Gamma(1-\sigma)^L}{\sigma^L \Gamma(1-\tilde{\sigma})^L} \frac{((\zeta+c)^{\sigma}-c^{\sigma})^L}{((\tilde{\zeta}+\tilde{c})^{\tilde{\sigma}}-\tilde{c}^{\tilde{\sigma}})^L} \left(\frac{\tilde{t}}{t}\right)^{a_t} e^{-b_t(\tilde{t}-t)} \prod_i \tilde{w}_i^{\sigma-\tilde{\sigma}} e^{-(\tilde{c}-c+\tilde{\zeta}-\zeta)\tilde{w}_i} \left(\frac{\tilde{\zeta}}{\zeta}\right)^{u_i} \\
= & \frac{\tilde{\sigma}^L \Gamma(1-\sigma)^L}{\sigma^L \Gamma(1-\tilde{\sigma})^L} \frac{((\zeta+c)^{\sigma}-c^{\sigma})^L}{((\tilde{\zeta}+\tilde{c})^{\tilde{\sigma}}-\tilde{c}^{\tilde{\sigma}})^L} \left(\frac{\tilde{t}}{t}\right)^{a_t} e^{-b_t(\tilde{t}-t)} \left(\prod_i \tilde{w}_i\right)^{\sigma-\tilde{\sigma}} e^{-(\tilde{c}-c+\tilde{\zeta}-\zeta)\sum_i \tilde{w}_i} \left(\frac{\tilde{\zeta}}{\zeta}\right)^{\sum_i u_i} \\
\log r = & L\left(\log\left(\frac{\tilde{\sigma}}{\sigma}\right) + \log\left(\frac{\Gamma(1-\sigma)}{\Gamma(1-\tilde{\sigma})}\right) + \log\left(\frac{((\zeta+c)^{\sigma}-c^{\sigma})}{((\tilde{\zeta}+\tilde{c})^{\tilde{\sigma}}-\tilde{c}^{\tilde{\sigma}})}\right)\right) + a_t \log\left(\frac{\tilde{t}}{t}\right) \\
& - b_t(\tilde{t}-t) + (\sigma-\tilde{\sigma})\left(\sum_i \log \tilde{w}_i\right) - (\tilde{c}-c+\tilde{\zeta}-\zeta)\left(\sum_i \tilde{w}_i\right) + \sum_i u_i \log\left(\frac{\tilde{\zeta}}{\zeta}\right) \\
= & \log \text{posterior}(\tilde{\sigma}, \tilde{c}, \tilde{t}|rest) - \log \text{posterior}(\sigma, c, t|rest) \\
& + \log q(\sigma, c, t|\tilde{\sigma}, \tilde{c}, \tilde{t}) - \log q(\tilde{\sigma}, \tilde{c}, \tilde{t}|\sigma, c, t)
\end{aligned}
$$

Metropolis-Hastings for x and γ

For x and γ we proceed in a way similar to t, σ, c. We update x and γ together, proposing a Metropolis-Hastings update:

$$
\begin{aligned}
\tilde{x}|x &\sim \text{trNormal}(0, x_{\max}, x, \sigma_x^2) \\
\tilde{\gamma}|\gamma &\sim \text{log Normal}(\log(\gamma), \sigma_\gamma^2)
\end{aligned}
$$

with trNormal(a, b, μ, σ^2) a truncated normal on the range $[a, b]$, with mean μ and standard deviation σ. We can compute the acceptance rate for x and γ by combining the likelihood, the priors and the proposals. We omit the computations, which are very similar to the ones for t, σ, c.

Gibbs sampler for $\tilde{w}, \bar{n}, u$

The full conditionals of $\bar{n}_{ij}, \tilde{w}_i$ and u_i are in closed form and easy to sample from, therefore we can construct the following Gibbs samplers:

- $\bar{n}_{ij}|rest \sim \text{trPoisson}(p_{ij}\tilde{w}_i\tilde{w}_j)$ for any i, j;
- $u_i|rest \overset{indep}{\sim} \text{trPoisson}(\zeta\tilde{w}_i)$ for any i;
- $\tilde{w}_i|rest, (\tilde{w}_j)_{j\neq i} \sim \text{Gamma}(-\sigma + u_i + \sum_{j\neq i} \bar{n}_{ij}, c + \zeta + \sum_{j\neq i} p_{ij}\tilde{w}_j)$ for any i.

Hamiltonian Monte Carlo for $\tilde{w}$

For the inference of the approximate sociabilities $\tilde{w}_i$, we propose also an Hamiltonian Monte Carlo procedure, that allows us to update all the weights together, instead of the sequential proposal of the Gibbs sampler. HMC is less subject to the presence of correlation among variables, which is the case instead for Gibbs samplers. To do so, we use as objective function the log posterior of the weights, of which we need to be able to compute the derivative. For numerical convenience, we take the derivative with respect to $\log \tilde{w}$.

$$
\begin{aligned}
\log p(\log \tilde{w}|rest) &= \log p(\tilde{w}|rest) + \sum_i \log \tilde{w}_i \\
\log p(\tilde{w}|rest) &\propto \sum_{i,j} [\bar{n}_{ij} \log(p_{ij}\tilde{w}_i\tilde{w}_j) - p_{ij}\tilde{w}_i\tilde{w}_j] \\
&\quad + \sum_i [(-1-\sigma)\log \tilde{w}_i - c\tilde{w}_i + u_i \log \tilde{w}_i - \zeta \tilde{w}_i] \\
\log p(\log \tilde{w}_i|rest) &\propto \log \tilde{w}_i + \log \tilde{w}_i \sum_j (\bar{n}_{ij} + \bar{n}_{ji}) - e^{\log \tilde{w}_i} \sum_{j \neq i} p_{ij}\tilde{w}_j - e^{2\log \tilde{w}_i} \\
&\quad + (u_i - 1 - \sigma)\log \tilde{w}_i - (\zeta + c)e^{\log \tilde{w}_i} \\
\frac{\partial \log p(\log \tilde{w}|rest)}{\partial \log \tilde{w}_i} &= m_i + u_i - \sigma - \tilde{w}_i(\zeta + \tau) - \tilde{w}_i \sum_j p_{ij}\tilde{w}_j
\end{aligned}
$$

with $m_i = \sum_j (\bar{n}_{ji} + \bar{n}_{ij})$. Fixing ϵ and R to be the leapfrog stepsize and number of steps, the algorithm iterates over the number of iterations according to these passages:

1. Sample the momentum Variables:

$$p_i \overset{iid}{\sim} \text{Normal}(0,1),\ i = 1, \dots, L$$

2. Simulate R steps of the discretized Hamiltonian: for every node i

$$
\begin{aligned}
\log \tilde{w}_i^{(0)} &= \log \tilde{w}_i \\
\tilde{p}_i^{(0)} &= p_i + \frac{\epsilon}{2} \frac{\partial(\log p(\log \tilde{w}|rest))}{\partial(\log \tilde{w}_i)}\Big|_{\log \tilde{w} = \log \tilde{w}^{(0)}} \\
\log \tilde{w}_i^{(r)} &= \log \tilde{w}_i^{(r-1)} + \epsilon \tilde{p}_i^{(r-1)} \quad r = 1, \dots, R-1 \\
\tilde{p}_i^{(r)} &= \tilde{p}_i^{(r-1)} + \epsilon \frac{\partial(\log p(\log \tilde{w}|rest))}{\partial(\log \tilde{w}_i)}\Big|_{\log \tilde{w} = \log \tilde{w}^{(r)}} \quad r = 1, \dots, R-1
\end{aligned}
$$

3. Set for every node i

$$
\begin{aligned}
\log w_i^* &= \log \tilde{w}_i^{(R-1)} + \epsilon \tilde{p}_i^{(R-1)} \\
\tilde{p}_i &= \tilde{p}_i^{(R-1)} + \frac{\epsilon}{2} \frac{\partial(\log p(\log \tilde{w}|rest))}{\partial(\log \tilde{w}_i)}\Big|_{\log \tilde{w} = \log w^*}
\end{aligned}
$$

4. Accept w^* with probability $\min(1, r_w)$, where

$$\log r_w = \log \left[\frac{p(\log w^*|rest)}{p(\log \tilde{w}|rest)} \times \frac{p(\tilde{p})}{p(p)} \right]$$

$$
\begin{aligned}
&= \sum_{i,j} \left[\bar{n}_{ij} (\log w_i^* + \log w_j^* - \log \tilde{w}_i - \log \tilde{w}_j) - p_{ij} (w_i^* w_j^* - \tilde{w}_i \tilde{w}_j) \right] \\
&\quad + \sum_i \left[(u_i - \sigma)(\log w_i^* - \log \tilde{w}_i) - (c + \zeta)(w_i^* - \tilde{w}_i) \right] \\
&\quad - \frac{1}{2} \sum_i (\tilde{p}_i^2 - p_i^2) = \\
&= \sum_i \left[(m_i + u_i - \sigma)(\log w_i^* - \log \tilde{w}_i) - (c + \zeta)(w_i^* - \tilde{w}_i) \right] \\
&\quad - \sum_{ij} [p_{ij} (w_i^* w_j^* - \tilde{w}_i \tilde{w}_j)] - \frac{1}{2} \sum_i (\tilde{p}_i^2 - p_i^2)
\end{aligned}
$$

References

Aalen, O. (1992). Modelling heterogeneity in survival analysis by the compound Poisson distribution. *The Annals of Applied Probability*, 951–972.

Aldous, D. (1997). Brownian excursions, critical random graphs and the multiplicative coalescent. *The Annals of Probability*, 812–854.

Aldous, D. (2009). More uses of exchangeability: representations of complex random structures. *arXiv preprint arXiv:0909.4339*.

Aldous, D. J. (1981). Representations for partially exchangeable arrays of random variables. *Journal of Multivariate Analysis 11*(4), 581–598.

Ayed, F., J. Lee, and F. Caron (2019). Beyond the chinese restaurant and pitman-yor processes: Statistical models with double power-law behavior. *arXiv preprint arXiv:1902.04714*.

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

Barthélemy, M. (2011). Spatial networks. *Physics reports 499*(1-3), 1–101.

Berger, N., C. Borgs, J. T. Chayes, and A. Saberi (2014). Asymptotic behavior and distributional limits of preferential attachment graphs. *The Annals of Probability 42*(1), 1–40.

Bianconi, G., P. Pin, and M. Marsili (2009). Assessing the relevance of node features for network structure. *Proceedings of the National Academy of Sciences 106*(28), 11433–11438.

Bingham, N. H., C. M. Goldie, and J. L. Teugels (1987). *Regular variation*, Volume 27. Cambridge university press.

Boguñá, M., F. Papadopoulos, and D. Krioukov (2010, sep). Sustaining the internet with hyperbolic mapping. *Nature Communications 1*(1).

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

Bollobás, B. (2001). *Random graphs*, Volume 73. Cambridge University Press.

Bollobás, B. and O. Riordan (2009). Metrics for sparse graphs. In S. Huczynska, J. Mitchell, and C. Roney-Dougal (Eds.), *Surveys in combinatorics*, Volume 365 of *London Mathematical Society Lecture Note Series*, pp. 211–287. arXiv:0708.1919: Cambridge University Press.

Borgs, C., J. Chayes, H. Cohn, and Y. Zhao (2019). An l^p theory of sparse graph convergence i: Limits, sparse random graph models, and power law distributions. *Transactions of the American Mathematical Society 372*(5), 3019–3062.

Borgs, C., J. T. Chayes, H. Cohn, and N. Holden (2016). Sparse exchangeable graphs and their limits via graphon processes. *arXiv preprint arXiv:1601.07134*.

Borgs, C., J. T. Chayes, and L. Lovász (2010). Moments of two-variable functions and the uniqueness of graph limits. *Geometric And Functional Analysis 19*(6), 1597–1619.

Bringmann, K., R. Keusch, and J. Lengler (2019). Geometric inhomogeneous random graphs. *Theoretical Computer Science 760*, 35–54.

Brix, A. (1999). Generalized gamma measures and shot-noise Cox processes. *Advances in Applied Probability 31*(4), 929–953.

Broido, A. D. and A. Clauset (2019). Scale-free networks are rare. *Nature communications 10*(1), 1–10.

Caldarelli, G. (2007). *Scale-free networks: complex webs in nature and technology*. Oxford University Press.

Caron, F. and E. Fox (2017). Sparse graphs using exchangeable random measures. *Journal of the Royal Statistical Society B 79*, 1–44. Part 5.

Caron, F., F. Panero, and J. Rousseau (2020). On sparsity and power-law and clustering properties of graphex processes. *arXiv preprint arXiv:1708.03120*.

Chung, F. and L. Lu (2002). The average distances in random graphs with given expected degrees. *Proceedings of the National Academy of Sciences 99*(25), 15879–15882.

Cohen, R. and S. Havlin (2010). *Complex networks: structure, robustness and function*. Cambridge university press.

Daley, D. J. and D. Vere-Jones (2008). *An Introduction to the Theory of Point Processes. Volume II: General Theory and Structure* (second ed.). Springer.

Dalmau, J. and M. Salvi (2019). Scale-free percolation in continuum space: quenched degree and clustering coefficient.

Deijfen, M., R. van der Hofstad, and G. Hooghiemstra (2013). Scale-free percolation. In *Annales de l'IHP Probabilités et statistiques*, Volume 49.

Deprez, P. and M. V. Wüthrich (2018, jul). Scale-free percolation in continuum space. *Communications in Mathematics and Statistics 7*(3), 269–308.

Dorogovtsev, S. N., S. N. Dorogovtsev, and J. F. Mendes (2003). *Evolution of networks: From biological nets to the Internet and WWW*. Oxford university press.

Durrett, R. (2007). *Random graph dynamics*. Cambridge university press.

Erdös, P. and A. Rényi (1959). On random graphs. *Publicationes Mathematicae 6*, 290–297.

Fienberg, S. E. (2012). A brief history of statistical models for network analysis and open challenges. *Journal of Computational and Graphical Statistics 21*(4), 825–839.

Geman, S. and D. Geman (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence* (6), 721–741.

Gilbert, E. (1961). Random plane networks. *Journal of the Society for Industrial and Applied Mathematics 9*(4), 533–543.

Haggett, P. and j. a. Chorley, Richard J. (1969). *Network analysis in geography*. London : Edward Arnold. Includes indexes.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika 57*(1), 97–109.

Hoff, P. D., A. E. Raftery, and M. S. Handcock (2002). Latent space approaches to social network analysis. *Journal of the american Statistical association 97*(460), 1090–1098.

Hougaard, P. (1986). Survival models for heterogeneous populations derived from stable distributions. *Biometrika 73*(2), 387–396.

James, L. F. (2002). Poisson process partition calculus with applications to exchangeable models and bayesian nonparametrics. *arXiv preprint math/0205093*.

Kingman, J. F. C. (1967). Completely random measures. *Pacific Journal of Mathematics 21*(1), 59–78.

Kingman, J. F. C. (1993). *Poisson processes*, Volume 3. Oxford University Press, USA.

Komjáthy, J. and B. Lodewijks (2019). Explosion in weighted hyperbolic random graphs and geometric inhomogeneous random graphs. *Stochastic Processes and their Applications*.

Krioukov, D., F. Papadopoulos, M. Kitsak, A. Vahdat, and M. Boguná (2010). Hyperbolic geometry of complex networks. *Physical Review E 82*(3), 036106.

Lee, J., L. F. James, and S. Choi (2016). Finite-dimensional bfry priors and variational bayesian inference for power law models. *Advances in Neural Information Processing Systems 29*.

Lee, M.-L. T. and G. A. Whitmore (1993). Stochastic processes directed by randomized time. *Journal of applied probability*, 302–314.

Lee, J., M. X. C. F. (2019). A unified construction for series representations and finite approximations of completely random measures.

Li, W. and X. Cai (2004, 05). Statistical analysis of airport network of china. *Physical review. E, Statistical, nonlinear, and soft matter physics 69*, 046106.

Lijoi, A., R. H. Mena, and I. Prünster (2007). Controlling the reinforcement in Bayesian nonparametric mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 69*(4), 715–740.

Lijoi, A. and I. Prünster (2003). On a normalized random measure with independent increments relevant to Bayesian nonparametric inference. In *Proceedings of the 13th European Young Statisticians Meeting*, pp. 123–134. Bernoulli Society.

Lovász, L. and B. Szegedy (2006). Limits of dense graph sequences. *Journal of Combinatorial Theory, Series B 96*(6), 933–957.

Neal, R. M. (2011). MCMC using Hamiltonian dynamics. In S. Brooks, A. Gelman, G. Jones, and X.-L. Meng (Eds.), *Handbook of Markov Chain Monte Carlo*, Volume 2. Chapman & Hall / CRC Press.

Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM review*, 167–256.

Newman, M. E. J. (2010). *Networks: An Introduction*. Oxford University Press.

Norros, I. and H. Reittu (2006). On a conditionally Poissonian graph process. *Advances in Applied Probability 38*(1), 59–75.

Orbanz, P. (2010). Conjugate projective limits.

Orbanz, P. and D. M. Roy (2015). Bayesian models of graphs, arrays and other exchangeable random structures. *IEEE Transactions on Pattern Analysis and Machine Intelligence 37*(2), 437–461.

Paleari, S., R. Redondi, and P. Malighetti (2010, March). A comparative study of airport connectivity in China, Europe and US: Which network provides the best service to passengers? *Transportation Research Part E: Logistics and Transportation Review 46*(2), 198–210.

Penrose, M. (2003). *Random geometric graphs*, Volume 5. Oxford University Press.

Rastelli, R., F. Maire, and N. Friel (2018). Computationally efficient inference for latent position network models. *arXiv preprint arXiv:1804.02274*.

Seshadri, M., S. Machiraju, A. Sridharan, J. Bolot, C. Faloutsos, and J. Leskove (2008). Mobile call graphs: Beyond power-law and lognormal distributions. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pp. 596–604. ACM.

Spencer, N. A. and C. R. Shalizi (2017). Projective, sparse, and learnable latent position network models.

Stoyan, D., W. S. Kendall, S. N. Chiu, and J. Mecke (2013). *Stochastic geometry and its applications*. John Wiley & Sons.

Swendsen, R. H. and J.-S. Wang (1986). Replica monte carlo simulation of spin-glasses. *Physical review letters 57*(21), 2607.

Van Der Hofstad, R. (2016). *Random Graphs and Complex Networks: Volume 1*, Volume 43. Cambridge university press.

Veitch, V. and D. M. Roy (2015). The class of random graphs arising from exchangeable random measures. *arXiv preprint arXiv:1512.03099*.

Voitalov, I., P. van der Hoorn, R. van der Hofstad, and D. Krioukov (2019). Scale-free networks well done. *Physical Review Research 1*(3), 033034.

Zhang, R. and C. M. De Sa (2019). Poisson-minibatching for gibbs sampling with convergence rate guarantees. *Advances in Neural Information Processing Systems 32*.

Zhang, Z. Q., F. C. Pu, and B. Z. Li (1983, feb). Long-range percolation in one dimension. *Journal of Physics A: Mathematical and General 16*(3), L85–L89.

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

Title of Paper	**Sparse spatial random graphs**
Publication Status	□Published □ Accepted for Publication □Submitted for Publication ☑ Unpublished and unsubmitted work written in a manuscript style
Publication Details	Joint work with Prof. François Caron (University of Oxford) and Prof. Judith Rousseau (University of Oxford)

Student Confirmation

Student Name:	Francesca Panero		
Contribution to the Paper	I am first author of this paper. I studied the asymptotic behaviour of the model, proving the results on sparsity, degree distribution and clustering coefficient. I implemented the sampling algorithm and proposed and implemented the MCMC for approximate posterior inference. I run the experiments on simulated data. I have carried out the study of the positioning of our model in the broader context of the spatial networks literature.		
Signature		Date	19/04/2022

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title:	Professor François Caron		
Supervisor comments			
Signature		Date	19/04/2022

This completed form should be included in the thesis, at the end of the relevant chapter.

Chapter 3

On sparsity, power-law and clustering properties of graphex processes

Submitted to the Advances in Applied Probability journal.

On sparsity, power-law and clustering properties of graphex processes

François Caron, Francesca Panero and Judith Rousseau

Department of Statistics, University of Oxford

This paper investigates properties of the class of graphs based on exchangeable point processes. We provide asymptotic expressions for the number of edges, number of nodes and degree distributions, identifying four regimes: (i) a dense regime, (ii) a sparse almost dense regime, (iii) a sparse regime with power-law behaviour, and (iv) an almost extremely sparse regime. We show that, under mild assumptions, both the global and local clustering coefficients converge to constants which may or may not be the same. We also derive a central limit theorem for subgraph counts and for the number of nodes. Finally, we propose a class of models within this framework where one can separately control the latent structure and the global sparsity/power-law properties of the graph.

MSC 2010 subject classifications: Primary 05C80, 60F15, 60G55.
Keywords: networks, sparsity, Poisson processes, community structure, power-law, generalised graphon, transitivity, subgraph counts.

1. Introduction

The ubiquitous availability of large, structured network data in various scientific areas ranging from biology to social sciences has been a driving force in the development of statistical network models (Kolaczyk, 2009; Newman, 2010). Vertex-exchangeable random graphs, also known as W-random graphs or graphon models (Hoover, 1979; Aldous, 1981; Lovász and Szegedy, 2006; Diaconis and Janson, 2008) offer in particular a flexible and tractable class of random graph models. It includes many models, such as the stochastic block-model (Nowicki and Snijders, 2001), as special cases. Various parametric and nonparametric model-based approaches (Palla et al., 2010; Lloyd et al., 2012; Latouche and Robin, 2016), or nonparametric estimation procedures (Wolfe and Olhede, 2013; Chatterjee, 2015; Gao et al., 2015) have been developed within this framework. Although very flexible, it is known that vertex-exchangeable random graphs are dense (Lovász and Szegedy, 2006; Orbanz and Roy, 2015), that is the number of edges scales quadratically with the number of nodes; this property is considered unrealistic for many real-world networks. To achieve sparsity, rescaled graphon models have been proposed in the literature (Bollobás and Riordan, 2009; Bickel and Chen, 2009; Bickel et al., 2011; Wolfe and Olhede, 2013). While these models can capture sparsity, they are not projective; additionally, standard rescaled graphon models cannot simultaneously capture sparsity and a clustering coefficient bounded away from 0 (see Section 5).

These limitations are overcome by another line of works initiated by Caron and Fox (2017), Veitch and Roy (2015) and Borgs et al. (2018). They showed that, by modeling the graph as an ex-

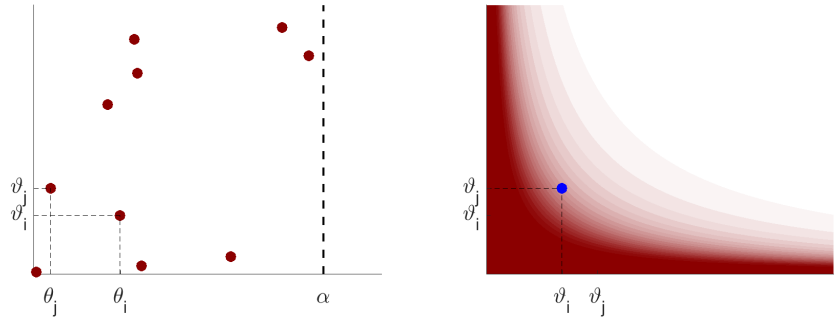


Figure 1. Illustration of the graph model based on exchangeable point processes. (left) A unit-rate Poisson process (θ_i, ϑ_i), $i \in \mathbb{N}$ on $(0, \alpha] \times \mathbb{R}_+$. (right) For each pair $i \leq j$, set $Z_{ij} = Z_{ji} = 1$ with probability $W(\vartheta_i, \vartheta_j)$. Here, W is indicated by the red shading (darker shading indicates higher value). Similar to Figure 5 in (Caron and Fox, 2017).

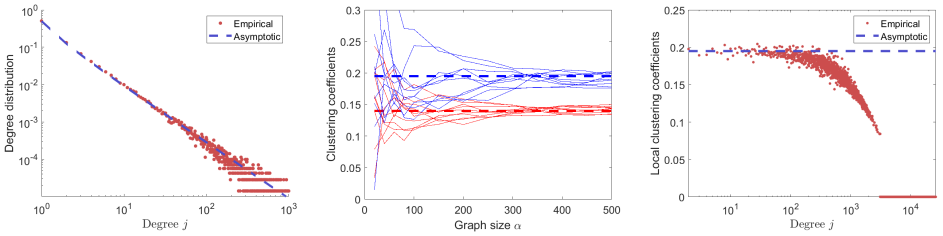
changeable point process, the classical vertex-exchangeable/graphon framework can be naturally extended to the sparse regime, while preserving its flexibility and tractability. In such a representation, introduced by Caron and Fox (2017), nodes are embedded at some location $\theta_i \in \mathbb{R}_+$, and the set of edges is represented by a point process on the plane

$$\sum_{i,j} Z_{ij} \delta_{(\theta_i, \theta_j)} \tag{1}$$

where $Z_{ij} = Z_{ji}$ is a binary variable indicating if there is an edge between node θ_i and node θ_j. Finite-size graphs are obtained by restricting the point process (1) to points (θ_i, θ_j) such that $\theta_i, \theta_j \leq \alpha$, with α a positive parameter controlling the size of the graph. Focusing on a particular construction as a case study, Caron and Fox (2017) showed that one can obtain sparse and exchangeable graphs within this framework; they also pointed out that exchangeable random measures admit a representation theorem due to Kallenberg (1990), giving a general construction for such graph models. Herlau et al. (2016), Todeschini et al. (2020) developed sparse graph models with (overlapping) community structure within this framework. Veitch and Roy (2015) and Borgs et al. (2018) showed how such construction naturally generalizes the dense exchangeable graphon framework to the sparse regime, and analysed some of the properties of the associated class of random graphs, called *graphex processes*[1]; further properties were derived by Janson (2016, 2017), Veitch and Roy (2019) and Borgs et al. (2019). Following the notations of Veitch and Roy (2015), and ignoring additional terms corresponding to stars and isolated edges, the graph is then parameterised by a symmetric measurable function $W : \mathbb{R}_+^2 \to [0, 1]$, where for each $i \leq j$,

$$Z_{ij} \mid (\theta_k, \vartheta_k)_{k=1,2,\ldots} \sim \text{Bernoulli}\{W(\vartheta_i, \vartheta_j)\}, \tag{2}$$

[1] Veitch and Roy (2015) introduced the term *graphex*. In the same paper, they referred to the class of random graphs as *Kallenberg exchangeable graphs*, but the term graphex processes is now more commonly used.



(a) Power-law Degree distribution (b) Average local and global clustering coefficients (c) Local clustering coefficient per degree

Figure 2. Illustration of some of the asymptotic results developed in this paper, applied to the generalised graphon model defined by Equations (39) and (44) with $\sigma_0 = 0.2$ and $\tau_0 = 2$. (a) Empirical degree distribution for a graph of size $\alpha = 1000$ (red) and asymptotic degree distribution (dashed blue, see Corollary 5). (b) Average local (blue) and global (red) clustering coefficients for 10 graphs of growing sizes. Limit values are represented by dashed lines (see Propositions 10 and 11). (c) Local clustering coefficient for nodes of a given degree j, for a graph of size $\alpha = 1000$. The limit value is represented by a dashed line (see Proposition 11).

where $(\theta_k, \vartheta_k)_{k=1,2,\ldots}$ is a unit-rate Poisson process on $\mathbb{R}_+^2$. See Figure 1 for an illustration of the model construction. The function W is a natural generalisation of the graphon for dense exchangeable graphs (Veitch and Roy, 2015; Borgs et al., 2018) and we refer to it as the graphon function.

This paper investigates asymptotic properties of the general class of graphs based on exchangeable point processes defined by Equations (1) and (2). Our findings can be summarised as follows.

(i) We relate the sparsity and power-law properties of the graph to the tail behaviour of the marginal of the graphon function W, identifying four regimes: a) a dense regime, b) a sparse (almost dense) regime without power-law behaviour, c) a sparse regime with power-law behaviour, and d) an almost extremely sparse regime. In the sparse, power-law regime, the power-law exponent is in the range $(1, 2)$.

(ii) We derive the asymptotic properties of the global and local clustering coefficients, two standard measures of the transitivity of the graph.

(iii) We give a central limit theorem for subgraph counts and for the number of nodes in the graph.

(iv) We introduce a parametrisation that allows to model separately the global sparsity structure and other local properties such as community structure. Such a framework enables us to sparsify any dense graphon model, and to characterise its sparsity properties.

(v) We show that the results apply to a wide range of sparse and dense graphex processes, including the models studied by Caron and Fox (2017), Herlau et al. (2016) and Todeschini et al. (2020).

Some of the asymptotic results are illustrated in Figure 2 for a specific graphex process in the sparse, power-law regime.

The article is organised as follows. In Section 2 we give the notations and the main Assump-

tions. In Section 3, we derive the asymptotic results for the number of nodes, degree distribution and clustering coefficients. In Section 4, we derive central limit theorems for subgraphs and for the number of nodes. Section 5 discusses related work. In Section 6 we provide specific examples of sparse and dense graphs and show how to apply the results of the previous section to those models. In Section 7 we describe a generic construction for graphs with local/global structure and adapt some results of Section 3 to this setting. Most of the proofs are given in the main text, with some longer proofs in the Appendix, together with some technical lemma and background material. Other more technical proofs are given in a Supplementary Material (Caron et al., 2020).

Throughout the document, we use the notations $X_\alpha \sim Y_\alpha$ and $X_\alpha = o(Y_\alpha)$ respectively for $X_\alpha/Y_\alpha \to 1$ and $X_\alpha/Y_\alpha \to 0$. Both notations $X_\alpha \lesssim Y_\alpha$ and $X_\alpha = O(Y_\alpha)$ are used for $\limsup X_\alpha/Y_\alpha < \infty$. The notation $X_\alpha \asymp Y_\alpha$ means both $X_\alpha \lesssim Y_\alpha$ and $Y_\alpha \lesssim X_\alpha$ hold. All unspecified limits are when α tends to infinity. When X_α and/or Y_α are random quantities, the asymptotic relation is meant to hold almost surely.

2. Notations and Assumptions

2.1. Notations

Let $M = \sum_i \delta_{(\theta_i,\vartheta_i)}$ be a unit-rate Poisson random measure on $(0,+\infty)^2$ and $W : [0,+\infty)^2 \to [0,1]$ a symmetric measurable function such that $\lim_{x\to\infty} W(x,x)$ and $\lim_{x\to 0} W(x,x)$ both exist [2] and

$$0 < \overline{W} = \int_{\mathbb{R}_+^2} W(x,y)dxdy < \infty, \quad \int_0^\infty W(x,x)dx < \infty, \tag{3}$$

Let $(U_{ij})_{i,j\in\mathbb{N}^2}$ be a symmetric array of independent random variables, with $U_{ij} \sim U(0,1)$ if $i \leq j$ and $U_{ij} = U_{ji}$ for $i > j$. Let $Z_{ij} = \mathbb{1}_{U_{ij}\leq W(\vartheta_i,\vartheta_j)}$ be a binary random variable indicating if there is a link between i and j, where $\mathbb{1}_A$ denotes the indicator function.

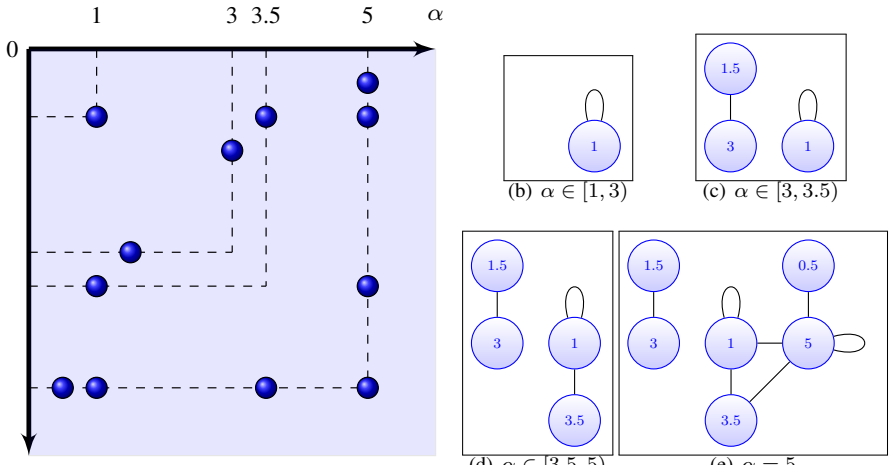
Restrictions of the point process $\sum_{ij} Z_{ij}\delta_{(\theta_i,\theta_j)}$ to squares $[0,\alpha]^2$ then define a growing family of random graphs $(\mathcal{G}_\alpha)_{\alpha\geq 0}$, called a graphex process, where $\mathcal{G}_\alpha = (\mathcal{V}_\alpha, \mathcal{E}_\alpha)$ denotes a graph of size $\alpha \geq 0$ with vertex set $\mathcal{V}_\alpha$ and edge set $\mathcal{E}_\alpha$, defined by

$$\mathcal{V}_\alpha = \{\theta_i \mid \theta_i \leq \alpha \text{ and } \exists \theta_k \leq \alpha \text{ s.t. } Z_{ik} = 1\} \tag{4}$$

$$\mathcal{E}_\alpha = \{\{\theta_i,\theta_j\} \mid \theta_i,\theta_j \leq \alpha \text{ and } Z_{ij} = 1\}. \tag{5}$$

The connection between the point process and graphex process is illustrated in Figure 3. The conditions (3) are sufficient (though not necessary) conditions for $|\mathcal{E}_\alpha|$ (hence $|\mathcal{V}_\alpha|$) to be almost surely finite, and the graphex process well defined (Veitch and Roy, 2015, Theorem 4.9). Note crucially that the graphs $\mathcal{G}_\alpha$ have no isolated vertices (that is, no vertices of degree 0), and that the number of nodes $|\mathcal{V}_\alpha|$ and edges $|\mathcal{E}_\alpha|$ are both random variables.

[2] By (3), this implies $\lim_{x\to\infty} W(x,x) = 0$.



(a) Point process $\sum_{ij} Z_{ij}\delta_{(\theta_i,\theta_j)}$ (b) $\alpha \in [1,3)$ (c) $\alpha \in [3,3.5)$ (d) $\alpha \in [3.5,5)$ (e) $\alpha = 5$

Figure 3. Illustration of the connection between the point process on the plane and the graphex process. (a) Point process $\sum_{ij} Z_{ij}\delta_{(\theta_i,\theta_j)}$ on the plane. (b-e) Associated graphs $\mathcal{G}_\alpha$ for (b) $\alpha \in [1,3)$, (c) $\alpha \in [3,3.5)$, (d) $\alpha \in [3.5,5)$ and (e) $\alpha = 5$. Note that the graph is empty for $\alpha < 1$.

We now define a number of summary statistics of the graph $\mathcal{G}_\alpha$. For $i \geq 1$, let

$$D_{\alpha,i} = \sum_k Z_{ik} \mathbb{1}_{\theta_k \leq \alpha}.$$

If $\theta_i \in \mathcal{V}_\alpha$, then $D_{\alpha,i} \geq 1$ corresponds to the degree of the node θ_i in the graph $\mathcal{G}_\alpha$ of size α; otherwise $D_{\alpha,i} = 0$. Let $N_\alpha = |\mathcal{V}_\alpha|$ and $N_{\alpha,j}$ be the number of nodes and the number of nodes of degree j, $j \geq 1$ respectively,

$$N_\alpha = \sum_i \mathbb{1}_{\theta_i \leq \alpha} \mathbb{1}_{D_{\alpha,i} \geq 1}, \quad N_{\alpha,j} = \sum_i \mathbb{1}_{\theta_i \leq \alpha} \mathbb{1}_{D_{\alpha,i} = j} \tag{6}$$

and $N_\alpha^{(e)} = |\mathcal{E}_\alpha|$ the number of edges

$$N_\alpha^{(e)} = \frac{1}{2} \sum_{i \neq j} Z_{ij} \mathbb{1}_{\theta_i \leq \alpha} \mathbb{1}_{\theta_j \leq \alpha} + \sum_i Z_{ii} \mathbb{1}_{\theta_i \leq \alpha}. \tag{7}$$

For $i \geq 1$, let

$$T_{\alpha,i} = \frac{1}{2} \sum_{j,k | j \neq k \neq i} Z_{ij} Z_{jk} Z_{ik} \mathbb{1}_{\theta_i \leq \alpha} \mathbb{1}_{\theta_j \leq \alpha} \mathbb{1}_{\theta_k \leq \alpha}. \tag{8}$$

If $\theta_i \in \mathcal{V}_\alpha$, $T_{\alpha,i}$ corresponds to the number of triangles containing node θ_i in the graph $\mathcal{G}_\alpha$, otherwise $T_{\alpha,i} = 0$. Let

$$T_\alpha = \frac{1}{3} \sum_i T_{\alpha,i} = \frac{1}{6} \sum_{i \neq j \neq k} Z_{ij} Z_{jk} Z_{ik} \mathbb{1}_{\theta_i \leq \alpha} \mathbb{1}_{\theta_j \leq \alpha} \mathbb{1}_{\theta_k \leq \alpha} \tag{9}$$

denote the total number of triangles and

$$A_\alpha = \sum_i \frac{D_{\alpha,i}(D_{\alpha,i}-1)}{2} = \frac{1}{2}\sum_{i\neq j\neq k} Z_{ij}Z_{jk}\mathbb{1}_{\theta_i\leq\alpha}\mathbb{1}_{\theta_j\leq\alpha}\mathbb{1}_{\theta_k\leq\alpha} \tag{10}$$

the total number of adjacent edges in the graph $\mathcal{G}_\alpha$. The global clustering coefficient, also known as the transitivity coefficient, is defined as

$$C_\alpha^{(g)} = \frac{3T_\alpha}{A_\alpha} \tag{11}$$

if $A_\alpha \geq 1$ and 0 otherwise. The global clustering coefficient counts the proportion of closed connected triplets over all the connected triplets, or equivalently the fraction of pairs of nodes connected to the same node that are themselves connected, and is a standard measure of the transitivity of a network (Newman, 2010, Section 7.9). Another measure of the transitivity of the graph is the local clustering coefficient. For any degree $j \geq 2$, define

$$C_{\alpha,j}^{(\ell)} = \frac{2}{j(j-1)N_{\alpha,j}}\sum_i T_{\alpha,i}\mathbb{1}_{D_{\alpha,i}=j} \tag{12}$$

if $N_{\alpha,j} \geq 1$ and 0 otherwise. $C_{\alpha,j}^{(\ell)}$ corresponds to the proportion of pairs of neighbours of nodes of degree j that are connected. The average local clustering coefficient is obtained by

$$\overline{C}_\alpha^{(\ell)} = \frac{1}{N_\alpha - N_{\alpha,1}}\sum_{j\geq 2} N_{\alpha,j}C_{\alpha,j}^{(\ell)} \tag{13}$$

if $N_\alpha - N_{\alpha,1} \geq 1$ and $\overline{C}_\alpha^{(\ell)} = 0$ otherwise.

2.2. Assumptions

We will make use of the following three assumptions. Assumption 1 characterises the behaviour of the small degree nodes. Assumption 2 is a technical assumption to obtain the almost sure results. Assumption 3 characterises the behaviour of large degree nodes.

A central quantity of interest in the analysis of the asymptotic properties of graphex processes is the marginal generalised graphon function $\mu : (0,\infty) \to \mathbb{R}_+$, defined for $x > 0$ by

$$\mu(x) = \int_0^\infty W(x,y)dy. \tag{14}$$

The integrability of the generalised graphon W implies that μ is integrable. Ignoring loops (self-edges), the expected number of connections of a node with parameter ϑ is proportional to $\mu(\vartheta)$. Therefore, assuming μ is monotone decreasing, its behaviour at infinity controls the small degree nodes, while its behaviour at 0 controls the large degree nodes.

For mathematical convenience, it will be easier to work with the generalised inverse μ^{-1} of μ. The behaviour at 0 of μ^{-1} then controls the small degree nodes, while the behaviour of μ^{-1} at infinity controls large degree nodes.

The following assumption characterises the behaviour of μ at infinity or, equivalently, of μ^{-1} at 0. We require μ^{-1} to behave approximately as a power function $x^{-\sigma}$ around 0, for some $\sigma \in [0,1]$. This behaviour, known as regular variation, has been extensively studied (see, e.g., Bingham et al. (1987)) and we provide some background on it in Appendix C.

Assumption 1 *Assume μ is non-increasing, with generalised inverse $\mu^{-1}(x) = \inf\{y > 0 \mid \mu(y) \leq x\}$, such that*

$$\mu^{-1}(x) \sim \ell(1/x)x^{-\sigma} \text{ as } x \to 0 \tag{15}$$

where $\sigma \in [0,1]$ and ℓ is a slowly varying function at infinity: for all $c > 0$, $\lim_{t\to\infty} \ell(ct)/\ell(t) = 1$.

Examples of slowly varying functions ℓ include functions converging to a strictly positive constant, or powers of logarithms. Note that Assumption 1 implies that, for $\sigma \in (0,1)$, $\mu(t) \sim \overline{\ell}(t)t^{-1/\sigma}$ as $t \to \infty$ for some slowly varying function $\overline{\ell}$. We can differentiate four cases, as it will be formally derived in Corollary 5.

(i) Dense case : $\sigma = 0$ and $\lim_{t\to\infty} \ell(t) < \infty$. In this case, $\lim_{x\to 0} \mu^{-1}(x) < \infty$, hence μ has bounded support. The other three cases are all sparse cases.

(ii) Almost dense case: $\sigma = 0$ and $\lim_{t\to\infty} \ell(t) = \infty$. In this case μ has full support and super-polynomially decaying tails.

(iii) Sparse case with power law : $\sigma \in (0,1)$. In this case μ has full support and polynomially decaying tails (up to a slowly varying function).

(iv) Very sparse case: $\sigma = 1$. In this case μ has full support and very light tails. In order for μ^{-1} (and hence W) to be integrable, we need ℓ to go to zero sufficiently fast.

Now define, for $x, y > 0$

$$\nu(x,y) = \int_0^\infty W(x,z)W(y,z)dz. \tag{16}$$

The expected number of common neighbours of nodes with parameters $(\vartheta_1, \vartheta_2)$ is proportional to $\nu(\vartheta_1, \vartheta_2)$.

The following assumption is a technical assumption needed in order to obtain the almost sure results on the number of nodes and degrees. Veitch and Roy (2015) made a similar assumption to obtain results in probability, see the discussion section for further details.

Assumption 2 *Assume that there exists $C_1, a > 0$ and $x_0 \geq 0$ such that for all $x, y > x_0$*

$$\nu(x,y) \leq C_1\mu(x)^a\mu(y)^a, \quad \mu(x_0) > 0, \quad \begin{cases} a > \max\left(\frac{1}{2}, \sigma\right) & \text{if } \sigma \in [0,1) \\ a = 1 & \text{if } \sigma = 1. \end{cases} \tag{17}$$

Remark 1 *Assumption 2 is trivially satisfied when the function W is separable $W(x,y) = \mu(x)\mu(y)/\overline{W}$. Assumptions 1 and 2 are also satisfied if*

$$W(x,y) = 1 - e^{-f(x)f(y)/\overline{f}} \tag{18}$$

for some positive, non-increasing, measurable function f with $\overline{f} = \int_0^\infty f(x)dx < \infty$ and generalised inverse f^{-1} verifying $f^{-1}(x) \sim \ell(1/x)x^{-\sigma}$ as x tends to 0. In this case, μ is monotone non-increasing. We have

$$\mu\{f^{-1}(x)\} = \int_0^\infty \{1 - e^{-xf(y)/\bar{f}}\}dy = x\int_0^\infty e^{-xu/\bar{f}} f^{-1}(u)/\bar{f}du \sim x$$

as x tends to 0 by dominated convergence. Hence $f\{\mu^{-1}(x)\} \sim x$ as x tends to 0 and $f^{-1}[f\{\mu^{-1}(x)\}] \sim \ell(1/x)x^{-\sigma}$. Assumption 2 follows from the inequality $W(x,y) \leq f(x)f(y)/\overline{f}$. Other examples are considered in Section 6.

The following assumption is used to characterise the asymptotic behaviour of large degree nodes.

Assumption 3 *Assume $\mu^{-1}(t) = \int_t^\infty f(x)dx$ where*

$$f(x) \sim \tau x^{-\tau-1}\ell_2(x) \text{ as } x \to \infty$$

where $\tau > 0$ and ℓ_2 is a slowly varying function.

Note that Assumption 3 implies that $\mu^{-1}(x) \sim x^{-\tau}\ell_2(x)$ as $x \to \infty$, and $\mu(t) \sim \bar{\ell}_2(t)t^{-1/\tau}$ as $t \to 0$ for some slowly varying function $\bar{\ell}_2$.

3. Asymptotic behaviour of various statistics of the graph

3.1. Asymptotic behaviour of the number of edges, number of nodes and degree distribution

In this section we characterise the almost sure and expected behaviour of the number of nodes N_α, number of edges $N_\alpha^{(e)}$ and number of nodes with j edges $N_{\alpha,j}$. These results allow us to provide precise statements about the sparsity of the graph and the asymptotic power-law properties of its degree distribution.

We first recall existing results on the asymptotic growth of the number of edges. The growth of the mean number of edges has been shown by Veitch and Roy (2015) and the almost sure convergence follows from (Borgs et al., 2018, Proposition 56).

Proposition 2 (Number of edges (Veitch and Roy, 2015; Borgs et al., 2018)) *As α goes to infinity, almost surely*

$$N_\alpha^{(e)} \sim E(N_\alpha^{(e)}) \sim \alpha^2\overline{W}/2. \tag{19}$$

The following two theorems provide a description of the asymptotic behaviour of the terms $N_\alpha, N_{\alpha,j}$ in expectation and almost surely.

Theorem 3 *For $\sigma \in [0,1]$, let ℓ_σ be slowly varying functions defined as*

$$\ell_1(t) = \int_t^\infty y^{-1}\ell(y)dy \quad \text{and} \quad \ell_\sigma(t) = \ell(t)\Gamma(1-\sigma) \text{ for } \sigma \in [0,1). \tag{20}$$

Under Assumption 1, for all $\sigma \in [0,1]$,

$$E(N_\alpha) \sim \alpha^{1+\sigma}\ell_\sigma(\alpha). \tag{21}$$

If $\sigma = 0$ *then for* $j \geq 1$

$$E(N_{\alpha,j}) = o\{\alpha\ell(\alpha)\}.$$

If $\sigma \in (0,1)$ *then for* $j \geq 1$

$$E(N_{\alpha,j}) \sim \frac{\sigma\Gamma(j-\sigma)}{j!}\alpha^{1+\sigma}\ell(\alpha)$$

Finally, if $\sigma = 1$,

$$E(N_{\alpha,j}) \sim \begin{cases} \alpha^2\ell_1(\alpha) & j = 1 \\ \frac{\alpha^2}{j(j-1)}\ell(\alpha) & j \geq 2 \end{cases}$$

Theorem 3 follows rather directly from asymptotic properties of regularly varying functions (Gnedin et al., 2007), recalled in Lemma B.2 and B.3 in the Appendix. Details of the proof are given in Appendix A.1.

Veitch and Roy (2015) have shown that, under Assumption 2 with $a = 1$, we have, in probability,

$$N_\alpha \sim E(N_\alpha), \quad \sum_{k\geq j} N_{\alpha,k} \sim E\left(\sum_{k\geq j} N_{\alpha,k}\right) \quad \text{for } j \geq 1.$$

The next theorem shows that the asymptotic equivalence holds almost surely under the weaker Assumption 2. Additionally, combining these results with Theorem 3 allows us to characterise the almost sure asymptotic behaviour of the number of nodes and number of nodes of a given degree. The proof of Theorem 4 is given in Section 3.2.

Theorem 4 *Under Assumptions 1 and 2, we have almost surely as* α *tends to infinity*

$$N_\alpha \sim E(N_\alpha), \quad \sum_{k\geq j} N_{\alpha,k} \sim E\left(\sum_{k\geq j} N_{\alpha,k}\right) \quad \textit{for } j \geq 1. \tag{22}$$

Combining this with Theorem 3, we obtain that, for all $\sigma \in [0,1]$,

$$N_\alpha \sim \alpha^{1+\sigma}\ell_\sigma(\alpha).$$

Moreover, for $j \geq 1$, *if* $\sigma = 0$ *then* $N_{\alpha,j} = o\{\alpha\ell(\alpha)\}$, *while if* $0 < \sigma < 1$

$$N_{\alpha,j} \sim \frac{\sigma\Gamma(j-\sigma)}{j!}\alpha^{1+\sigma}\ell(\alpha).$$

If $\sigma = 1$, $N_{\alpha,1} \sim \alpha^2\ell_1(\alpha)$ *and for all* $j \geq 2$ *we also have,* $N_{\alpha,j} = o\{\alpha^2\ell_1(\alpha)\}$.

The following result is a corollary of Theorem 4 which shows how the parameter σ relates to the sparsity and power-law properties of the graphs. We denote $\ell^{\#}$ the de Bruijn conjugate (see definition C.3 in the Appendix) of the slowly varying function ℓ.

Corollary 5 (Sparsity and power-law degree distribution) *Assume Assumptions 1 and 2. For $\sigma \in [0,1]$, almost surely as α tends to infinity,*

$$N_\alpha^{(e)} \sim \frac{\overline{W}}{2} N_\alpha^{2/(1+\sigma)} \ell_\sigma^*(N_\alpha), \quad \ell_\sigma^*(y) = \left[\left\{ \ell_\sigma^{1/(1+\sigma)}(y^{1/1+\sigma}) \right\}^{\#} \right]^2.$$

$\ell_\sigma^*(y)$ is slow varying and the graph is dense if $\sigma = 0$ and $\lim_{t\to\infty} \ell(t) = C < \infty$, as $N_\alpha^{(e)}/N_\alpha^2 \to C^2\overline{W}/2$ almost surely. Otherwise, if $\sigma > 0$ or $\sigma = 0$ and $\lim_t \ell(t) = \infty$, the graph is sparse, as $N_\alpha^{(e)}/N_\alpha^2 \to 0$. Additionally, for $\sigma \in [0,1)$, for any $j = 1, 2, \ldots$,

$$\frac{N_{\alpha,j}}{N_\alpha} \to \frac{\sigma\Gamma(j-\sigma)}{j!\Gamma(1-\sigma)} \tag{23}$$

almost surely. If $\sigma > 0$, this corresponds to a degree distribution with a power-law behaviour as, for j large

$$\frac{\sigma\Gamma(j-\sigma)}{j!\Gamma(1-\sigma)} \sim \frac{\sigma}{\Gamma(1-\sigma)j^{1+\sigma}}.$$

For $\sigma = 1$, $N_{\alpha,1}/N_\alpha \to 1$ and $N_{\alpha,j}/N_\alpha \to 0$ for $j \geq 2$, hence the nodes of degree 1 dominate in the graph.

Remark 6 *If $\sigma = 0$ and $\lim_{t\to\infty} \ell(t) = \infty$, the graph is almost dense, that is $N_\alpha^{(e)}/N_\alpha^2 \to 0$ and $N_\alpha^{(e)}/N_\alpha^{2-\epsilon} \to \infty$ for any $\epsilon > 0$. If $\sigma = 1$, the graph is almost extremely sparse (Bollobás and Riordan, 2009), as $N_\alpha^{(e)}/N_\alpha \to \infty$ and $N_\alpha^{(e)}/N_\alpha^{1+\epsilon} \to 0$ for any $\epsilon > 0$.*

The above results are important in terms of modelling aspects, since they allow a precise description of the degrees and number of edges as a function of the number of nodes. They can also be used to conduct inference on the parameters of the statistical network model, since the behaviour of most estimators will depend heavily on the behaviour of $N_\alpha, N_\alpha^{(e)}$ and possibly $N_{\alpha,j}$. For instance the following naive estimator[3] of σ

$$\hat{\sigma} = \frac{2\log N_\alpha}{\log N_\alpha^{(e)}} - 1 \tag{24}$$

is almost surely consistent. Indeed under Assumptions 1 and 2, using Theorems 2 and 4, we have almost surely $N_\alpha^2 \sim \alpha^{2+2\sigma}\ell_\sigma(\alpha)^2$ and $N_\alpha^{(e)} \sim \alpha^2\overline{W}/2$. Hence

$$\log \frac{N_\alpha^2}{N_\alpha^{(e)}} \sim 2\sigma\log(\alpha) + \log\{\ell_\sigma(\alpha)^2 2/\overline{W}\}$$

[3] Following an earlier version of the present paper, Naulet et al. (2017) proposed an alternative estimator for σ, with better statistical properties.

and the result follows as $\log \ell_\sigma(\alpha)/\log \alpha \to 0$.

All the above results depend on the behaviour of small degree nodes. It is also of interest to look at, for α fixed, the number of nodes of degree j for large j. We show in the next proposition that this is controlled by the high degree nodes.

Proposition 7 (Power-law for high degree nodes) *Assume that Assumption 3 holds for some $\tau > 0$ and some slowly varying function ℓ_2. Then, for fixed graph size α*

$$E(N_{\alpha,j}) \sim \frac{\alpha^{\tau+1}\tau\ell_2(j)}{j^{1+\tau}} \text{ as } j \to \infty.$$

This corresponds to a power-law behaviour with exponent $1+\tau$.

Proof. Under Assumption 3, we have $\mu^{-1}(t) = \int_t^\infty f(x)dx$ with

$$f(x) \sim \tau x^{-\tau-1}\ell_2(x) \text{ as } x \to \infty$$

where $\tau > 0$ and ℓ_2 is a slowly varying function. It implies $\lim_{t\to 0}\mu(t) = \infty$. From (Veitch and Roy, 2015, Theorem 5.5), we have

$$\begin{aligned} E(N_{\alpha,j}) &= \alpha\int_0^\infty (1 - W(\vartheta,\vartheta))e^{-\alpha\mu(\vartheta)}\frac{(\alpha\mu(\vartheta))^j}{j!}d\vartheta + \alpha\int_0^\infty e^{-\alpha\mu(\vartheta)}W(\vartheta,\vartheta)\frac{(\alpha\mu(\vartheta))^{j-1}}{(j-1)!}d\vartheta \\ &= \alpha\int_0^\infty (1 - W(\mu^{-1}(x),\mu^{-1}(x)))e^{-\alpha x}\frac{(\alpha x)^j}{j!}f(x)dx \\ &\quad + \alpha\int_0^\infty e^{-\alpha x}W(\mu^{-1}(x),\mu^{-1}(x))\frac{(\alpha x)^{j-1}}{(j-1)!}f(x)dx \end{aligned}$$

Note that $\lim_{x\to\infty} W(\mu^{-1}(x),\mu^{-1}(x)) \in [0,1]$ exists. Using Corollary B.6 (Willmot, 1990, Theorem 2.1.) in the Appendix, we obtain that, for fixed α,

$$E(N_{\alpha,j}) \sim \frac{\alpha^{\tau+1}\tau\ell_2(j)}{j^{1+\tau}} \text{ as } j \to \infty.$$

■

3.2. Proof of Theorem 4

The proof follows similarly to that of (Veitch and Roy, 2015, Theorem 3.1), by bounding the variance. Veitch and Roy (2015) showed that $\mathrm{var}(N_\alpha) = o(E(N_\alpha)^2)$ and $\mathrm{var}(N_{\alpha,j}) = o(E(N_{\alpha,j})^2)$ and use this result to prove that (22) holds in probability; we need a slightly tighter bound on the variances to obtain the almost sure convergence. This is stated in the next two Propositions.

Proposition 8 *Let N_α be the number of nodes. We have*

$$\begin{aligned}\mathrm{var}(N_\alpha) = E(N_\alpha) + 2\alpha^2 \int_{\mathbb{R}_+} \mu(x)\{1 - W(x,x)\}e^{-\alpha\mu(x)}dx \\ + \alpha^2 \int_{\mathbb{R}_+^2} \{1 - W(x,y)\}\{1 - W(x,x)\}\{1 - W(y,y)\} \\ \left\{e^{\alpha\nu(x,y)} - 1 + W(x,y)\right\} e^{-\alpha\mu(x)-\alpha\mu(y)}dxdy. \end{aligned} \tag{25}$$

Under Assumptions 1 and 2, with $\sigma \in [0,1]$, slowly varying function ℓ and positive scalar a satisfying (17), we have

$$\mathrm{var}(N_\alpha) = O\{\alpha^{3+2\sigma-2a}\ell_\sigma(\alpha)^2\}. \tag{26}$$

where the slowly varying functions ℓ_σ are defined in Equation (20). Additionally, under Assumptions 1 and 4, we have, for any $\sigma \in [0,1]$ and any slowly varying function ℓ

$$\mathrm{var}(N_\alpha) \asymp \alpha^{1+2\sigma}\ell_\sigma^2(\alpha). \tag{27}$$

The proof of Proposition 8 is given in Section S1.1 in the Supplementary Material (Caron et al., 2020). Proposition 8 and Theorem 3 imply in particular that, under Assumptions 1 and 2,

$$\mathrm{var}(N_\alpha) = O\{E(N_\alpha)^2\alpha^{-\kappa}\}$$

for some $\kappa > 0$. N_α is a positive, monotone increasing stochastic process. Using Lemma B.1 in the Appendix, we obtain that $N_\alpha \sim E(N_\alpha)$ almost surely as α tends to ∞.

Proposition 9 *Let $N_{\alpha,j}$ be the number of nodes of degree j. Then, under Assumptions 1 and 2, with $\sigma \in [0,1]$, slowly varying function ℓ and positive scalar a satisfying (17), we have*

$$\mathrm{var}(N_{\alpha,j}) = O\{\alpha^{3+2\sigma-2a}\ell_\sigma(\alpha)^2\}.$$

where the slowly varying functions ℓ_σ are defined in Equation (20). In the case $\sigma = 0$ and $a = 1$, we have the stronger result

$$\mathrm{var}(N_{\alpha,j}) = o\{\alpha\ell(\alpha)^2\}.$$

The proof of Proposition 9 is given in Section S1.2 in the Supplementary Material (Caron et al., 2020). Define $\widetilde{N}_{\alpha,j} = \sum_{k\geq j} N_{\alpha,k}$, the number of nodes of degree at least j. Note that $\widetilde{N}_{\alpha,j}$ is a positive, monotone increasing stochastic process in α, with $\widetilde{N}_{\alpha,j} = N_\alpha - \sum_{k=1}^{j-1} N_{\alpha,k}$. We then have that, using Cauchy-Schwarz and Jensen's inequalities

$$E(\widetilde{N}_{\alpha,j}) = E(N_\alpha) - \sum_{k=1}^{j-1} E(N_{\alpha,k}), \quad \mathrm{var}(\widetilde{N}_{\alpha,j}) \leq j\left\{\mathrm{var}(N_\alpha) + \sum_{k=1}^{j-1}\mathrm{var}(N_{\alpha,k})\right\}.$$

Consider first the case $\sigma \in [0,1)$. Since Theorem 3 implies, for $j \geq 2$, $\alpha^{1+\sigma}\ell(\alpha) \lesssim E(\widetilde{N}_{\alpha,j})$ as α goes to infinity, using Propositions 8 and 9, we obtain $\mathrm{var}(\widetilde{N}_{\alpha,j}) = O\{\alpha^{-\tau}E(\widetilde{N}_{\alpha,j})^2\}$ for some $\tau > 0$. Combined with Lemma B.1, it leads to $\widetilde{N}_{\alpha,j} \sim E(\widetilde{N}_{\alpha,j})$ almost surely as α goes to infinity.

The almost sure results for $N_{\alpha,j}$ then follow from the fact that, for all $j \geq 2$, $E(\widetilde{N}_{\alpha,j}) \asymp E(N_\alpha)$ if $\sigma \in (0,1)$, $E(\widetilde{N}_{\alpha,j}) \sim E(N_\alpha)$ if $\sigma = 0$ and $E(\widetilde{N}_{\alpha,j}) = o\{E(N_\alpha)\}$ if $\sigma = 1$.

3.3. Asymptotic behaviour of the clustering coefficients

The following Proposition is a direct corollary of (Borgs et al., 2018, Proposition 56) who showed the almost sure convergence of subgraph counts in graphex processes.

Proposition 10 (Global clustering coefficient (Borgs et al., 2018)) *Assume* $\int_0^\infty \mu(x)^2 dx < \infty$. *Recall that* T_α *and* A_α *are respectively the number of triangles and number of adjacent edges in the graph of size* α. *We have*

$$T_\alpha \sim E(T_\alpha) = \frac{\alpha^3}{6} \int_{\mathbb{R}_+^3} W(x,y)W(x,z)W(y,z)dxdydz,$$
$$A_\alpha \sim E(A_\alpha) = \frac{\alpha^3}{2} \int_0^\infty \mu(x)^2 dx$$

almost surely as $\alpha \to \infty$. *Therefore, if* $\int_0^\infty \mu(x)^2 dx > 0$, *the global clustering coefficient defined in Equation* (11) *converges to a constant*

$$C_\alpha^{(g)} \to \frac{\int_{\mathbb{R}_+^3} W(x,y)W(x,z)W(y,z)dxdydz}{\int_0^\infty \mu(x)^2 dx} \text{ almost surely as } \alpha \to \infty.$$

Note that if μ is monotone decreasing, as $\overline{W} < \infty$, we necessarily have $\int_a^\infty \mu(x)^2 dx < \infty$ for any $a > 0$. Hence the condition $\int_0^\infty \mu(x)^2 dx < \infty$ in Proposition 10 requires additional assumptions on the behaviour of μ at 0 (or equivalently the behaviour of μ^{-1} at ∞), which drives the behaviour of large degree nodes. If the graph is dense, μ is bounded and thus $\int_0^\infty \mu(x)^2 dx < \infty$.

Proposition 11 (Local clustering coefficient) *Assume Assumptions 1 and 2 hold with* $\sigma \in (0,1)$. *Assume additionally that*

$$\lim_{x\to\infty} \frac{\int_{\mathbb{R}_+^2} W(x,y)W(x,z)W(y,z)dydz}{\mu(x)^2} \to b \tag{28}$$

for some $b \in [0,1]$. *Then the local clustering coefficients converge in probabiltiy as* $\alpha \to \infty$:

$$C_{\alpha,j}^{(\ell)} \to b \quad \forall j \geq 2.$$

If $b > 0$, *the above result holds almost surely, and the average local clustering coefficient satisfies*

$$\lim_{\alpha\to\infty} \overline{C}_\alpha^{(\ell)} \to b, \quad \textit{almost surely}.$$

In general,

$$\lim_{x\to\infty} \frac{1}{\mu(x)^2} \int W(x,y)W(x,z)W(y,z)dydz \neq \frac{\int W(x,y)W(x,z)W(y,z)dxdydz}{\int \mu(x)^2 dx}$$

and the global clustering and local clustering coefficients converge to different limits. A notable exception is the separable case where $W(x,y) = \mu(x)\mu(y)/\overline{W}$, since in this case

$$\int W(x,y)W(x,z)W(y,z)dydz = \overline{W}^{-3}\mu(x)^2\left(\int \mu(y)^2 dy\right)^2, \quad b = \frac{\left(\int \mu(y)^2 dy\right)^2}{\overline{W}^3}$$

and

$$\int W(x,y)W(x,z)W(y,z)dydzdx = \overline{W}^{-3}\left(\int \mu(y)^2 dy\right)^3.$$

Sketch of the proof. Full details are given in Appendix A.2, and we only give here a sketch of the proof, which is similar to that of Theorem 4. We have

$$C^{(\ell)}_{\alpha,j} = \frac{2R_{\alpha,j}}{j(j-1)N_{\alpha,j}}, \quad \text{where} \quad R_{\alpha,j} = \sum_i T_{\alpha,i}\mathbb{1}_{D_{\alpha,i}=j}$$

$R_{\alpha,j}$ corresponds to the number of triangles having a node of degree j as a vertex, where triangles having$k \leq 3$ degree-j nodes as vertices are counted k times.

We obtain an asymptotic expression for $E(R_{\alpha,j})$, and show that $\mathrm{var}(R_{\alpha,j}) = O(\alpha^{1-2a}[E(R_{\alpha,j})]^2)$. We then prove that $R_{\alpha,j}/E(R_{\alpha,j})$ goes to 1 almost surely. The latter is obtained by proving that $R_{\alpha,j}$ is nearly monotonic increasing by constructing an increasing sequence α_n going to infinity such that $E(R_{\alpha_n,j})/E(R_{\alpha_{n+1},j})$ goes to 1 and such that for all $\alpha \in (\alpha_n, \alpha_{n+1})$

$$R_{\alpha_n,j} - \tilde{R}_{n,j} \leq R_{\alpha,j} \leq R_{\alpha_{n+1},j} + \tilde{R}_{n,j}, \quad \tilde{R}_{n,j} = o_p(E(R_{\alpha_n,j})).$$

Roughly speaking $\tilde{R}_{n,j}$ corresponds to the sum of the number of triangles from i, over the set i such that $D_{n,i} \leq j$ and i has at least one connection with some i' such that $\theta_{i'} \in (\alpha_n, \alpha_{n+1})$. The result for the local clustering coefficient then follows from Toeplitz lemma (see e.g. (Loève, 1977, p. 250)).

4. Central limit theorems

We now present central limit theorems (CLT) for subgraph counts (number of edges, triangles, etc.) and for the number of nodes N_α. Subgraph counts can be expressed as U-statistics of Poisson random measures (up to an asymptotically negligible term). A CLT then follows rather directly from CLT on U-statistics of Poisson random measures (Reitzner and Schulte, 2013).

Obtaining a CLT for quantities like N_α is more challenging, since these cannot be reduced to U-statistics. We prove in this Section the CLT for N_α and we separate the dense and sparse cases because the techniques of the respective proofs are very different. The proof of the sparse case requires additional assumptions and is much more involved. We believe that the same technique of proof can be used for other quantities of interest, such as the number $N_{\alpha,j}$ of nodes of degree j; with more tedious computations.

4.1. CLT for subgraph counts

4.1.1. Statement of the result

Let F be a given subgraph, which has neither isolated vertices nor loops. Denote $|F|$ the number of nodes, $\{1, \cdots, |F|\}$ the set of vertices and $e(F)$ the set of edges. Let $N_\alpha^{(F)}$ be the number of subgraphs F in the graph $\mathcal{G}_\alpha$:

$$N_\alpha^{(F)} = k_{(F)} \sum_{(v_1,\cdots,v_{|F|})}^{\neq} \prod_{(i,j)\in e(F)} Z_{v_i,v_j} \mathbb{1}_{\theta_{v_i}\leq\alpha} \mathbb{1}_{\theta_{v_j}\leq\alpha},$$

where $k_{(F)}$ is a constant accounting for the multiple counts of F, that we can omit in the rest of the discussion since it does not depend on α. Note that this statistics covers the number of edges (excluding loops) if $|F| = 2$ and the number of triangles if $|F| = 3$ and $e(F) = \{(1,2),(1,3),(2,3)\}$. It is known in the graph literature as the number of injective adjacency maps from the vertex set of F to the vertex set of $\mathcal{G}_\alpha$, see (Borgs et al., 2018, Section 2.5).

Proposition 12 *Let F be a subgraph without self edges nor isolated vertices. Assume that $\int_0^\infty \mu(x)^{2|F|-2}dx < \infty$. Then*

$$\frac{N_\alpha^{(F)} - E\left(N_\alpha^{(F)}\right)}{\sqrt{\text{var}\left(N_\alpha^{(F)}\right)}} \to \mathcal{N}(0,1), \tag{29}$$

as α goes to infinity, where

$$E(N_\alpha^{(F)}) = k_{(F)}\alpha^{|F|} \int_{\mathbb{R}_+^{|F|}} \prod_{(i,j)\in e(F)} W(x_i,x_j)dx_1\cdots dx_{|F|} < \infty \tag{30}$$

and

$$\text{var}(N_\alpha^{(F)}) \sim c_F \alpha^{2|F|-1}$$

for some positive constant c_F that depends only on F.

Remark 13 *If the graph is dense, μ is a bounded function with bounded support and therefore $\int_0^\infty \mu(x)^p dx < \infty$ for any p. In the sparse case, if μ is monotone, we necessarily have $\int_a^\infty \mu(x)^p dx < \infty$ for any $p > 1$. The condition $\int_0^\infty \mu(x)^{2|F|-2}dx < \infty$ therefore requires additional assumptions on the behaviour of μ at 0, which drives the behaviour of large degree nodes.*

4.1.2. Proof

The main idea of the proof is to use the decomposition

$$N_\alpha^{(F)} - E(N_\alpha^{(F)}) = E(N_\alpha^{(F)}|M) - E(N_\alpha^{(F)}) + N_\alpha^{(F)} - E(N_\alpha^{(F)}|M), \tag{31}$$

and to show that $E(N_\alpha^{(F)}|M)$ is a geometric U-Statistic of a Poisson process, for which CLT have been derived by Reitzner and Schulte (2013).

In this section, denote $K = |F| \geq 2$ the number of nodes of the subgraph F. The subgraph counts are

$$N_\alpha^{(F)} = k_{(F)} \sum_{(v_1,\cdots,v_K)}^{\neq} \left(\prod_{k=1}^{K} \mathbb{1}_{\theta_{v_k} \leq \alpha} \right) \frac{1}{|\mathbb{S}_K|} \sum_{\pi \in \mathbb{S}_K} \prod_{(i,j)\in e(F)} Z_{v_{\pi_i}, v_{\pi_j}}$$

where $\mathbb{S}_K$ denotes the set of permutations of $\{1, \ldots, K\}$.

Using the extended Slivnyak-Mecke theorem, we have

$$E(N_\alpha^{(F)}) = k_{(F)} \alpha^K \int_{\mathbb{R}_+^K} \prod_{(i,j)\in e(F)} W(x_i, x_j) dx_1 \cdots dx_K. \tag{32}$$

As $\int_0^\infty \mu(x)^{K-1} dx < \infty$, Lemma 62 in (Borgs et al., 2018) implies that $E(N_\alpha^{(F)}) < \infty$. For any $K \geq 2$, define the symmetric function

$$f(x_1, \ldots, x_K) = \frac{1}{|\mathbb{S}_K|} \sum_{\pi \in \mathbb{S}_K} \prod_{(i,j)\in e(F)} W(x_{\pi_i}, x_{\pi_j});$$

additionally, using condition (3) and $\int_0^\infty \mu(x)^{K-1} dx < \infty$, it satisfies $0 < \int_{\mathbb{R}_+^K} f(x_1, \ldots, x_K) dx_1 \ldots dx_K < \infty$.

We state the following useful lemma.

Lemma 14 *The function f satisfies for all $x_K \geq 0$*

$$g(x_K) := \int_{\mathbb{R}_+^{K-1}} f(x_1, \ldots, x_{K-1}, x_K) dx_1 \ldots dx_{K-1} \leq C_0 \max(\mu(x_K), \mu(x_K)^{K-1})$$

for some constant C_0.

Proof. Let $\pi \in \mathbb{S}_K$ and $r_K \in \{1, \ldots, K\}$ be such that $\pi_{r_K} = K$. Denote $S \subseteq \{1, \ldots, K-1\}$ the set of indices i such that $(i, r_K) \in e(F)$ and i has no other connections in F. Then

$$\begin{aligned} \int_{\mathbb{R}_+^{K-1}} \prod_{(i,j)\in e(F)} W(x_{\pi_i}, x_{\pi_j}) dx_1 \ldots dx_{K-1} &\leq C_1 \int_{\mathbb{R}_+^{|S|}} \left[\prod_{i \in S} W(x_{\pi_i}, x_K) dx_i \right] \\ &= C_1 \mu(x_K)^{|S|} \leq C_1 \max(\mu(x_K), \mu(x_K)^{K-1}) \end{aligned}$$

for some constant C_1. ■

It follows from Lemma 14 and from the fact that $\int_0^\infty \mu(x) dx < \infty$ that, if $\int_0^\infty \mu(x)^{2K-2} dx < \infty$, then

$$\int_0^\infty \left(\int_{\mathbb{R}_+^{K-1}} f(x_1, \ldots, x_{K-1}, y) dx_1 \ldots dx_{K-1} \right)^2 dy < \infty.$$

We are now ready to derive the asymptotic expression for the variance of $N_\alpha^{(F)}$. Using the extended Slivnyak-Mecke theorem again,

$$
\begin{aligned}
&E((N_\alpha^{(F)})^2) = E(E((N_\alpha^{(F)})^2 \mid M)) \\
&= k_{(F)}^2 E\left(\sum_{(v_1,\cdots,v_K,v'_1,\ldots,v'_K)}^{\neq} f(\vartheta_{v_1},\ldots,\vartheta_{v_K}) f(\vartheta_{v'_1},\ldots,\vartheta_{v'_K}) \prod_{k=1}^{K} \mathbb{1}_{\theta_{v_k}\le\alpha} \mathbb{1}_{\theta_{v'_k}\le\alpha} \right) \\
&+ k_{(F)}^2 K^2 E\left(\sum_{\substack{(v_1,\cdots,v_K,\\ v'_1,\ldots,v'_{K-1})}}^{\neq} f(\vartheta_{v_1},\ldots,\vartheta_{v_K}) f(\vartheta_{v'_1},\ldots,\vartheta_{v'_{K-1}},\vartheta_{v_K}) \mathbb{1}_{\theta_{v_K}\le\alpha} \prod_{k=1}^{K-1} \mathbb{1}_{\theta_{v_k}\le\alpha} \mathbb{1}_{\theta_{v'_k}\le\alpha} \right) \\
&+ O(\alpha^{2K-2}) \\
&= k_{(F)}^2 K^2 \alpha^{2K-1} \int_{\mathbb{R}_+^{2K-1}} f(x_1,\ldots,x_K) f(x'_1,\ldots,x'_{K-1},x_K) dx_1,\ldots dx_K dx'_1 \ldots dx'_{K-1} \\
&+ E(N_\alpha^{(F)})^2 + O(\alpha^{2K-2}).
\end{aligned}
$$

It follows that

$$\mathrm{var}(N_\alpha^{(F)}) \sim k_{(F)}^2 K^2 \alpha^{2K-1} \sigma_F^2$$

as α tends to infinity, where

$$\sigma_F^2 = \int_0^\infty \left(\int_{\mathbb{R}_+^{K-1}} f(x_1,\ldots,x_{K-1},y) dx_1 \ldots dx_{K-1} \right)^2 dy < \infty.$$

We now prove the CLT. The first term of the right-handside of Equation (31) takes the form

$$E(N_\alpha^{(F)}|M) = k_{(F)} \sum_{(v_1,\cdots,v_K)}^{\neq} f(\vartheta_{v_1},\ldots,\vartheta_{v_K}) \prod_{i=1}^{K} \mathbb{1}_{\theta_{v_i}\le\alpha}. \tag{33}$$

By the superposition property of Poisson random measures, we have

$$E(N_\alpha^{(F)}|M) \stackrel{d}{=} k_{(F)} \sum_{(v_1,\cdots,v_K)}^{\neq} f(\widetilde{\vartheta}_{v_1},\ldots,\widetilde{\vartheta}_{v_K}) \prod_{i=1}^{K} \mathbb{1}_{\widetilde{\theta}_{v_i}\le 1}$$

where the right-handside is a geometric U-statistic (Reitzner and Schulte, 2013, Definition 5.1) of the Poisson point process $\{(\widetilde{\theta}_i, \widetilde{\vartheta}_i)_{i\ge 1}\}$ with mean measure $\alpha d\widetilde{\theta} d\widetilde{\vartheta}$ on $[0,1] \times \mathbb{R}_+$. Theorem 5.2 in Reitzner and Schulte (2013) therefore implies that

$$\frac{E(N_\alpha^{(F)} \mid M) - E(N_\alpha^{(F)})}{\sqrt{\mathrm{var}(E(N_\alpha^{(F)} \mid M))}} \to \mathcal{N}(0,1) \tag{34}$$

where $\mathrm{var}(E(N_\alpha^{(F)} \mid M)) \sim \mathrm{var}(N_\alpha^{(F)}) \sim k_{(F)}^2 |F|^2 \alpha^{2|F|-1} \sigma_F^2$. One can show similarly (proof omitted) that $\mathrm{var}(N_\alpha^{(F)} - E(N_\alpha^{(F)} \mid M)) = o(\alpha^{2|F|-1})$. It follows from Equations (31), (34) and Chebyshev inequality that

$$\frac{N_\alpha^{(F)} - E(N_\alpha^{(F)})}{\sqrt{\mathrm{var}(N_\alpha^{(F)})}} \to \mathcal{N}(0,1)$$

as α tends to infinity.

4.2. CLT for N_α (dense case)

4.2.1. Statement of the result

In the dense case, μ has a bounded support. If it is monotone decreasing, then Assumption 1 is satisfied with $\sigma = 0$ and $\ell(t) = \sup\{x > 0 \mid \mu(x) > 0\}$ is constant. In this case a central limit theorem (CLT) applies, as described in the following theorem.

Theorem 15 (Dense case) *Assume that Assumption 1 holds with $\sigma = 0$ and $\ell(t) = C \in (0, \infty)$ where $C = \sup\{x > 0 \mid \mu(x) > 0\}$ (dense case). Also assume Assumption 2 holds with $a = 1$. Then*

$$\frac{N_\alpha - E(N_\alpha)}{\sqrt{\alpha C}} \to \mathcal{N}(0,1). \tag{35}$$

Moreover, $E(N_\alpha) = \alpha C - m_{\alpha,0}$ where

$$m_{\alpha,0} = \alpha \int_0^C e^{-\alpha\mu(x)}(1 - W(x,x))dx = o(\alpha). \tag{36}$$

$m_{\alpha,0}$ can be interpreted as the expected number of degree 0 nodes, and is finite in the dense case. $m_{\alpha,0}$ can either diverge or converge to a constant as α tends to infinity, as shown in the following examples.

Example 16 *Consider $\mu(x) = \mathbb{1}_{x\in[0,1]}$, $\mu(x) = (1-x)^2\mathbb{1}_{x\in[0,1]}$ and $\mu(x) = (1-x)^3\mathbb{1}_{x\in[0,1]}$. We respectively have $m_{\alpha,0} \to 0$, $m_{\alpha,0} \sim \frac{\sqrt{\pi}}{2}\alpha^{1/2}$ and $m_{\alpha,0} \sim \Gamma(4/3)\alpha^{2/3}$.*

The above CLT for N_α can be generalised to $\widetilde{N}_{\alpha,j} = \sum_{k\geq j} N_{\alpha,k}$, the number of nodes of degree at least j.

Theorem 17 *Assume that Assumption 1 holds with $\sigma = 0$ and $\ell(t) = C \in (0,\infty)$ where $C = \sup\{x > 0 \mid \mu(x) > 0\}$ (dense case). Also assume Assumption 2 holds with $a = 1$. Then, for any $j \geq 1$*

$$\frac{\widetilde{N}_{\alpha,j} - E(\widetilde{N}_{\alpha,j})}{\sqrt{\alpha C}} \to \mathcal{N}(0,1). \tag{37}$$

Moreover, $E(\widetilde{N}_{\alpha,1}) = E(N_\alpha) = \alpha C - m_{\alpha,0}$ and for $j \geq 2$, $E(\widetilde{N}_{\alpha,j}) = \alpha C - m_{\alpha,0} - \sum_{k=1}^{j-1} E(N_{\alpha,j})$ where $m_{\alpha,0}$ is defined in Equation (36) and $E(N_{\alpha,j})$ is defined in Equation (51). Note that $m_{\alpha,0} = o(\alpha)$ and for any $j \geq 1$, $E(N_{\alpha,j}) = o(\alpha)$.

4.2.2. Proof

For a point (θ, ϑ) such that $\vartheta > C$, its degree is necessarily equal to zero, as $\mu(\vartheta) = 0$. Write

$$N_\alpha = Q_\alpha - N_{\alpha,0}, \quad \text{where} \quad Q_\alpha = \sum_i \mathbb{1}_{\theta_i \leq \alpha} \mathbb{1}_{\vartheta_i \leq C};$$

Q_α is the total number of nodes i with $\theta_i \leq \alpha$ that could have a connection (hence such that $\mu(\vartheta_i) > 0$), and

$$N_{\alpha,0} = \sum_i \mathbb{1}_{\theta_i \leq \alpha} \mathbb{1}_{\vartheta_i \leq C} \mathbb{1}_{D_{\alpha,i}=0}$$

is the set of nodes i with degree 0, but for which $\theta_i \leq \alpha, \mu(\vartheta_i) > 0$. In the dense regime, both Q_α and $N_{\alpha,0}$ are almost surely finite. $(Q_\alpha)_{\alpha \geq 0}$ is a homogeneous Poisson process with rate C. By the law of large numbers, $Q_\alpha \sim \alpha C \sim N_\alpha$ almost surely as α tends to infinity. Using Campbell's theorem, the Slivnyak-Mecke formula, and monotone convergence, we have$E(N_{\alpha,0}) = \alpha \int_0^C (1 - W(x,x)) e^{-\alpha\mu(x)} dx = o(\alpha)$. We also have that

$$E(N_{\alpha,0}^2) - E(N_{\alpha,0}) = \alpha^2 \int_0^C \int_0^C (1 - W(x,x))(1 - W(y,y))(1 - W(x,y)) e^{-\alpha\mu(x) - \alpha\mu(y) + \alpha\nu(x,y)} dxdy.$$

Hence, using the inequality $e^x - 1 \leq xe^x$, we obtain

$$\begin{aligned}
\mathrm{var}(N_{\alpha,0}) &= \alpha^2 \int_0^C \int_0^C (1 - W(x,x))(1 - W(y,y))(1 - W(x,y)) e^{-\alpha\mu(x) - \alpha\mu(y) + \alpha\nu(x,y)} dxdy \\
&\quad - \alpha^2 \left(\int_0^C (1 - W(x,x)) e^{-\alpha\mu(x)} dx \right)^2 + E(N_{\alpha,0}) \\
&\leq E(N_{\alpha,0}) + \alpha^3 \int_0^C \int_0^C \nu(x,y) e^{-\alpha\mu(x) - \alpha\mu(y) + \alpha\nu(x,y)} dxdy.
\end{aligned}$$

Using Lemma B.7 in the Appendix and Assumption 2 with $a = 1$,

$$\int_0^C \int_0^C \nu(x,y) e^{-\alpha/2\mu(x) - \alpha/2\mu(y)} dxdy = o(\alpha^{-2}).$$

It follows that $\mathrm{var}(N_{\alpha,0}) = o(\alpha)$. This implies, using Chebyshev's inequality, the CLT for Poisson processes and Slutsky's theorem that

$$\frac{N_\alpha - E(N_\alpha)}{\sqrt{\alpha C}} = \frac{Q_\alpha - \alpha C}{\sqrt{\alpha C}} - \frac{N_{\alpha,0} - E(N_{\alpha,0})}{\sqrt{\alpha C}} \to \mathcal{N}(0,1).$$

This concludes the proof of Theorem 15. The proof of Theorem 17 follows similarly. Note that the case $j = 1$ in Theorem 17 corresponds to Theorem 15. For any $j \geq 2$, $\widetilde{N}_{\alpha,j} = Q_\alpha - N_{\alpha,0} - \sum_{k=1}^{j-1} N_{\alpha,k}$. We have, using Cauchy-Schwarz inequality and Proposition 9,

$$\mathrm{var}\left(N_{\alpha,0} + \sum_{k=1}^{j-1} N_{\alpha,k} \right) \leq j \left(\mathrm{var}(N_{\alpha,0}) + \sum_{k=1}^{j-1} \mathrm{var}(N_{\alpha,k}) \right) = o(\alpha)$$

This implies

$$\frac{\widetilde{N}_{\alpha,j} - E(\widetilde{N}_{\alpha,j})}{\sqrt{\alpha C}} = \frac{Q_\alpha - \alpha C}{\sqrt{\alpha C}} - \frac{N_{\alpha,0} + \sum_{k=1}^{j-1} N_{\alpha,k} - E(N_{\alpha,0} + \sum_{k=1}^{j-1} N_{\alpha,k})}{\sqrt{\alpha C}} \to \mathcal{N}(0,1).$$

4.3. CLT for N_α (sparse case)

4.3.1. Statement of the result

We now assume that we are in the sparse regime, that is μ has unbounded support. We make the following additional assumptions in order to prove the asymptotic normality. Both assumptions hold when W is separable, as well as in the model of Caron and Fox (2017) under some moment conditions (see Section 6.5).

Assumption 4 *Assume that there exists $0 < C_0 \leq C_1$ and $x_0 \geq 0$ such that for all $x, y > x_0$*

$$C_0\mu(x)\mu(y) \leq \nu(x,y) \leq C_1\mu(x)\mu(y).$$

Assumption 5 *Assume that for any $j \leq 6$, and any $(x_1, \ldots, x_j) \in \mathbb{R}_+^j$*

$$\int_0^\infty \prod_{i=1}^j W(x_i, y)dy \leq \prod_{i=1}^j L(x_i)\mu(x_i)$$

where L is a locally integrable, slowly varying function converging to a (strictly positive) constant, and such that

$$\int_0^\infty L(x)\mu(x)dx < \infty.$$

Obviously, Assumption 4 implies that Assumption 2 is satisfied with $a = 1$.

We now state the central limit theorem for N_α under the sparse regime. Recall that in this case, when Assumption 1 holds, we either have $\sigma = 0$ and $\ell(t) \to \infty$ or $\sigma \in (0, 1]$.

Theorem 18 (Sparse case) *Assume that μ has an unbounded support (sparse regime). Under Assumptions 1, 4 and 5, we have*

$$\frac{N_\alpha - E(N_\alpha)}{\sqrt{\mathrm{var}(N_\alpha)}} \to \mathcal{N}(0,1).$$

Remark 19 *As detailed in Proposition 8, under Assumptions 1 and 4, we have, for any $\sigma \in [0,1]$ and any slowly varying function ℓ, $\mathrm{var}(N_\alpha) \asymp \alpha^{1+2\sigma}\ell_\sigma^2(\alpha)$ where the slowly varying function ℓ_σ is defined in Equation (20).*

4.3.2. Proof

The proof uses the recent results of Last et al. (2016) on normal approximations of non-linear functions of a Poisson random measure. We have the decomposition

$$\begin{aligned} N_\alpha - E(N_\alpha) &= (N_\alpha - E(N_\alpha \mid M)) + (E(N_\alpha \mid M) - E(N_\alpha)) \\ &= (N_\alpha - E(N_\alpha \mid M)) + (M(h_\alpha) - E(N_\alpha)) + f_\alpha(M) \end{aligned}$$

where

$$f_\alpha(M) = \sum_i \mathbb{1}_{\theta_i \le \alpha} \left[(1 - W(\vartheta_i, \vartheta_i)) e^{-\alpha\mu(\vartheta_i)} - e^{-M(g_{\alpha,\vartheta_i})} \right]$$

is a nonlinear functional of the Poisson random measure M, and

$$M(h_\alpha) = \sum_i \mathbb{1}_{\theta_i \le \alpha} \left[1 - (1 - W(\vartheta_i, \vartheta_i)) e^{-\alpha\mu(\vartheta_i)} \right]$$

is a linear functional of M with $h_\alpha(\theta, \vartheta) = \mathbb{1}_{\theta \le \alpha} \left[1 - (1 - W(\vartheta, \vartheta)) e^{-\alpha\mu(\vartheta)} \right]$. Theorem 18 is a direct consequence of the following three propositions and of Slutsky's theorem.

Proposition 20 *Under Assumptions 1 and 4, we have*

$$N_\alpha - E(N_\alpha \mid M) = \begin{cases} O(\alpha^{1/2+\sigma/2} \ell_\sigma^{1/2}(\alpha)) & \text{if } \sigma \in [0, 1) \\ o(\alpha^{1/2} \ell^{1/2}(\alpha)) & \text{if } \sigma = 0 \end{cases} \quad \text{in probability}$$

hence

$$\frac{N_\alpha - E(N_\alpha \mid M)}{\sqrt{\mathrm{var}(N_\alpha)}} \to 0 \text{ in probability.}$$

Proposition 21 *Under Assumptions 1 and 4, we have*

$$M(h_\alpha) - E(N_\alpha) = O(\alpha^{1/2+\sigma/2} \ell^{1/2}(\alpha)) \text{ in probability}$$

hence, if μ has an unbounded support,

$$\frac{M(h_\alpha) - E(N_\alpha)}{\sqrt{\mathrm{var}(N_\alpha)}} \to 0 \text{ in probability.}$$

Proposition 22 *Assume μ has an unbounded support. Under Assumptions 1, 4 and 5, we have*

$$\frac{f_\alpha(M)}{\sqrt{\mathrm{var}(N_\alpha)}} \to \mathcal{N}(0, 1).$$

The above three propositions are proved in Section S2 of the Supplementary Material (Caron et al., 2020). The most challenging part is Proposition 22, where we use the results of (Last et al., 2016, Theorem 1.1) on the normal approximation of non-linear functionals of Poisson random measures. This requires the determination of tight bounds on integrals involving expectations of multiple moments of terms of the form

$$D_z F_\alpha = \frac{f_\alpha(M + \delta_z) - f_\alpha(M)}{\sqrt{\mathrm{var}(N_\alpha)}},$$
$$D_{z_1,z_2} F_\alpha = \frac{f_\alpha(M + \delta_{z_1} + \delta_{z_2})) - f_\alpha(M + \delta_{z_1}) - f_\alpha(M + \delta_{z_2}) + f_\alpha(M)}{\sqrt{\mathrm{var}(N_\alpha)}}.$$

The obtention of these bounds requires tedious calculations. The details are given in Section S2.3 of the Supplementary Material (Caron et al., 2020).

5. Related work and Discussion

Veitch and Roy (2015) proved that Equation (22) holds in probability, under slightly different assumptions: they assume that Assumption 2 holds with $a = 1$ and that μ is differentiable, with some conditions on the derivative, but do not make any assumption on the existence of σ or ℓ. We note that for all the examples considered in Section 6, Assumptions 1 and 2 are always satisfied, but Assumption 2 does not hold with $a = 1$ for the non-separable graphon function (38). Additionally, the differentiability condition does not hold for some standard graphon models such as the stochastic blockmodel. Borgs et al. (2018) proved, amongst other results, the almost sure convergence of the subgraph counts in graphex models (Theorem 56). For the subclass of graphon models defined by Equation (39), Caron and Fox (2017) provided a lower bound on the growth in the number of nodes, and therefore an upper bound on the sparsity rate, using assumptions of regular variation similar to (40). Applying the results derived in this Section, we show in Section 6.5 that the bound is tight, and we derive additional asymptotic properties for this particular class.

As mentioned in the introduction, another class of (non projective) models that can produce sparse graphs are sparse graphons (Bollobás and Riordan, 2009; Bickel and Chen, 2009; Bickel et al., 2011; Wolfe and Olhede, 2013). In particular, a number of authors considered the following sparse graphon model, where two nodes i and j in a graph of size n connect with probability $\rho_n W(U_i, U_j)$ where $W : [0,1]^2 \to [0,1]$ is the symmetric graphon function, measurable and symmetric and $\rho_n \to 0$. Although such model can capture sparsity, it has rather different properties compared to those of graphex models. For example, the global clustering coefficient for this sparse graphon model converges to 0, while the clustering coefficient converges to a positive constant, as shown in Proposition 10.

Also graphex processes include as a special case dense vertex-exchangeable random graphs (Hoover, 1979; Aldous, 1981; Lovász and Szegedy, 2006; Diaconis and Janson, 2008), that is models based on a graphon on $[0,1]$. They also include as a special case the class of graphon models over more general probability spaces (Bollobás et al., 2007); see (Borgs et al., 2018, p.21) for more details. Some other classes of graphs, such as geometric graphs arising from Poisson processes in different spaces (Penrose, 2003), cannot be cast in this framework.

6. Examples of sparse and dense models

We provide here some examples of the four different cases: dense, almost dense, sparse and almost extremely sparse. We also show that the results of the previous section apply to the particular model studied by Caron and Fox (2017).

6.1. Dense graph

Let us consider the graphon function

$$W(x,y) = (1-x)(1-y)\,\mathbb{1}_{x\leq 1}\,\mathbb{1}_{y\leq 1}$$

which has bounded support. The corresponding marginal graphon function $\mu(x) = \mathbb{1}_{x\leq 1}(1-x)/2$ has inverse $\mu^{-1}(x) = \ell(1/x)$ where $\ell(1/x) = (1-2x)\mathbb{1}_{x\leq 1/2}$ is slowly varying since $\ell(1/x) \to 1$. Assumptions 1 and 2 are satisfied, hence by Theorem 4 and Corollary 5

$$N_\alpha \sim \alpha, \quad N_\alpha^{(e)} \sim \alpha^2/8, \quad N_\alpha^{(e)} \sim N_\alpha^2/8, \quad \frac{N_{\alpha,j}}{N_\alpha} \to 0 \quad j \geq 1$$

almost surely as $\alpha \to \infty$. The function W is separable and $C_\alpha^{(g)} \to 4/9$.

6.2. Sparse, almost dense graph without power-law

Consider the graphon function, considered by Veitch and Roy (2015),

$$W(x,y) = e^{-x-y}$$

which has full support. The corresponding function $\mu(x) = e^{-x}$ has inverse $\mu^{-1}(x) = \ell(1/x) = \log(1/x)\mathbb{1}_{0<x<1}$, which is a slowly varying function. We have $\ell_0^*(x) = 1/\log(x)^2$. Assumptions 1 and 2 are satisfied and

$$N_\alpha \sim \alpha\log(\alpha), \quad N_\alpha^{(e)} \sim \alpha^2/2, \quad N_\alpha^{(e)} \sim \frac{N_\alpha^2}{2\log(N_\alpha)^2}, \quad \frac{N_{\alpha,j}}{N_\alpha} \to 0 \text{ for all } j = 1,2,\ldots$$

The function W is separable, and $C_\alpha^{(g)} \to 1/4$.

6.3. Sparse graphs with power-law

We consider two examples here, a separable and a non-separable one. Interestingly, while both examples have similar power-law behaviours regarding the degree distribution, the clustering properties are very different. In the first example, the local clustering coefficient converges to a strictly positive constant, while in the second example, it converges to 0.

Separable example. First, consider the function

$$W(x,y) = (x+1)^{-1/\sigma}(y+1)^{-1/\sigma}$$

with $\sigma \in (0,1)$. We have $\mu(x) = \sigma(x+1)^{-1/\sigma}/(1-\sigma)$, $\mu^{-1}(x) = x^{-\sigma}(1/\sigma-1)^{-\sigma} - 1$, $\ell(t) \sim (1/\sigma-1)^{-\sigma}$ and $\ell_\sigma^*(t) \sim \{(1/\sigma-1)^{-\sigma}\Gamma(1-\sigma)\}^{-2/(1+\sigma)}$. Assumptions 1 and 2 are satisfied. We have $N_\alpha \sim \alpha^{1+\sigma}\Gamma(1-\sigma)(1/\sigma-1)^{-\sigma}$, $N_\alpha^{(e)} \sim \alpha^2\sigma^2/\{2(1-\sigma)^2\}$ and

$$N_\alpha^{(e)} \sim \frac{\sigma^2\left\{\Gamma(1-\sigma)(\frac{1}{\sigma}-1)^{-\sigma}\right\}^{-\frac{2}{1+\sigma}}}{2(1-\sigma)^2} N_\alpha^{2/(1+\sigma)}, \quad \frac{N_{\alpha,j}}{N_\alpha} \to \frac{\sigma\Gamma(j-\sigma)}{j!\Gamma(1-\sigma)}, \quad j \geq 1.$$

The function is separable, and we obtain, for $\sigma \in (0,1)$

$$\lim_{\alpha\to\infty} C_\alpha^{(g)} = \left(\frac{1-\sigma}{2-\sigma}\right)^2 \text{ and } \lim_{\alpha\to\infty} C_{\alpha,j}^{(\ell)} = \left(\frac{1-\sigma}{2-\sigma}\right)^2 \text{ almost surely.}$$

Non-separable example. Consider now the non-separable function

$$W(x,y) = (x+y+1)^{-1/\sigma-1} \tag{38}$$

where $\sigma \in (0,1)$. We have $\mu(x) = \sigma(x+1)^{-1/\sigma}$, $\mu^{-1}(x) = \sigma^\sigma x^{-\sigma} - 1$, $\ell(t) \sim \sigma^\sigma$ and $\ell^*_\sigma(t) \sim \{\sigma^\sigma\Gamma(1-\sigma)\}^{-2/(1+\sigma)}$. Assumptions 1 and 2 are satisfied as for all $(x,y) \in \mathbb{R}^2_+$

$$W(x,y) \leq (x+1)^{-1/(2\sigma)-1/2}(y+1)^{-1/(2\sigma)-1/2} = \sigma^{-1-\sigma}\mu(x)^{\frac{1+\sigma}{2}}\mu(y)^{\frac{1+\sigma}{2}}.$$

We have $N_\alpha \sim \alpha^{1+\sigma}\Gamma(1-\sigma)\sigma^\sigma$, $N^{(e)}_\alpha \sim \alpha^2\sigma^2/\{2(1-\sigma)\}$ and

$$N^{(e)}_\alpha \sim \frac{\sigma^2\left[\Gamma(1-\sigma)\sigma^\sigma\right]^{-\frac{2}{1+\sigma}}}{2(1-\sigma)} N_\alpha^{2/(1+\sigma)}, \quad \frac{N_{\alpha,j}}{N_\alpha} \to \frac{\sigma\Gamma(j-\sigma)}{j!\Gamma(1-\sigma)}, \quad j \geq 1.$$

We have $\int \mu(x)^2dx = \frac{\sigma^3}{2-\sigma}$. There is no analytical expression for $\int W(x,y)W(y,z)W(x,z)dxdydz$, but this quantity can be evaluated numerically, and is non-zero, so the global clustering coefficient converges almost surely to a non-zero constant for any $\sigma \in (0,1)$. For the local clustering coefficient, we have $\mu(x)^2 \sim \sigma^2 x^{-2/\sigma}$ as $x \to \infty$ and

$$\int W(x,y)W(x,z)W(y,z)dydz \leq x^{-2/\sigma-2}\int (y+z+1)dydz = o(\mu(x)^2).$$

Hence the local clustering coefficients $C^{(\ell)}_{\alpha,j}$ converge in probability to 0 for all j.

6.4. Almost extremely sparse graph

Consider the function

$$W(x,y) = \frac{1}{(x+1)(1+\log(1+x))^2}\frac{1}{(y+1)(1+\log(1+y))^2}.$$

We have $\overline{W} = 1$ and $\mu(x) = (x+1)^{-1}(1+\log(1+x))^{-2}$ and, using properties of inverses of regularly varying functions, $\mu^{-1}(x) \sim x^{-1}\ell(1/x)$ as $x \to 0$, where $\ell(t) = \log(t)^{-2}$ is a slowly varying function. We have, for $t > 1$, $\ell_1(t) = \int_t^\infty x^{-1}\ell(x)dx = 1/\log(t)$ and $\ell^*_1(t) \sim \log(t)/2$. Assumptions 1 and 2 are satisfied, and almost surely

$$N^{(e)}_\alpha \sim \alpha^2/2, \quad N_\alpha \sim \frac{\alpha^2}{\log(\alpha)}, \quad N^{(e)}_\alpha \sim \frac{1}{4}N_\alpha\log(N_\alpha),$$
$$\frac{N_{\alpha,1}}{N_\alpha} \to 1, \quad \frac{N_{\alpha,j}}{N_\alpha} \to 0 \text{ for all } j \geq 2.$$

$\int \mu(x)^2dx = \frac{1}{6}(2+e\text{Ei}(-1)) \simeq 0.24$ where Ei is the exponential integral, hence $C^{(g)}_\alpha \to 0.0576$ almost surely.

6.5. Model of Caron and Fox (2017)

Caron and Fox (2017) studied a particular subclass of non-separable graphon models. This class is very flexible and allows to span the whole range of sparsity and power-law behaviours described in Section 3. As shown by Caron and Fox (2017), efficient Monte Carlo algorithms can be developed for estimating the parameters of this class of models. Additionally, (Borgs et al., 2019, Corollary 1.3) recently showed that this class is the limit of some sparse configuration models, providing further motivation for the study of their mathematical properties.

Let ρ be a Lévy measure on $(0, +\infty)$ and $\overline{\rho}(x) = \int_x^\infty \rho(dw)$ the corresponding tail Lévy intensity with generalised inverse $\overline{\rho}^{-1}(x) = \inf\{u > 0 | \overline{\rho}(u) < x\}$. Caron and Fox (2017) introduced the model defined by

$$W(x,y) = \begin{cases} 1 - e^{-2\overline{\rho}^{-1}(x)\overline{\rho}^{-1}(y)} & x \neq y \\ 1 - e^{-\{\overline{\rho}^{-1}(x)\}^2} & x = y \end{cases} . \tag{39}$$

$w = \overline{\rho}^{-1}(x)$ can be interpreted as the sociability of a node with parameter x. The larger this value, the more likely it is to connect to other nodes. The tail Lévy intensity $\overline{\rho}$ is a monotone decreasing function; its behaviour at 0 will control the low degree nodes while its behaviour at infinity will control the behaviour of high degree nodes.

The following proposition formalises this and shows how the results of Sections 3 and 4 apply to this model. Its proof is given in Section 6.6.

Proposition 23 *Consider the graphon function W defined by Equation (39) with Lévy measure ρ and tail Lévy intensity $\overline{\rho}$. Assume $m = \int_0^\infty w\rho(dw) < \infty$ and*

$$\overline{\rho}(x) \sim x^{-\sigma}\widetilde{\ell}(1/x) \text{ as } x \to 0 \tag{40}$$

for some $\sigma \in [0,1]$ and some slowly varying function $\widetilde{\ell}$. Then Equation (3) and Assumptions 1 and 2 hold, with $a = 1$ and $\ell(x) = (2m)^\sigma \widetilde{\ell}(x)$. Proposition 2, Theorems 3, 4 and Corollary 5 therefore hold. If $\int_0^\infty \psi(2w)^2\rho(dw) < \infty$, where $\psi(t) = \int(1 - e^{-wt})\rho(dw)$ is the Laplace exponent, then the global clustering coefficient converges almost surely

$$\lim_{\alpha\to\infty} C_\alpha^{(g)} = \frac{\int_{\mathbb{R}_+^3}(1-e^{-2xy})(1-e^{-2xz})(1-e^{-2yz})\rho(dx)\rho(dy)\rho(dz)}{\int_0^\infty \psi(2w)^2\rho(dw)}$$

and when $\sigma \in (0,1)$, Proposition 11 holds and for any $j \geq 2$

$$\lim_{\alpha\to\infty} C_{\alpha,j}^{(\ell)} = \lim_{\alpha\to\infty} \overline{C}_\alpha^{(\ell)} = 1 - \frac{\int_{\mathbb{R}_+^2} yze^{-2yz}\rho(dy)\rho(dz)}{m^2},$$

almost surely. For a given subgraph F, the CLT for the number of such subgraphs (Proposition 12) holds if $\int \psi(2\overline{\rho}^{-1}(x))^{2|F|-2}dx < \infty$. Under Assumption 1, this condition always holds if $\sigma = 0$; for $\sigma \in (0,1]$, it holds if $\overline{\rho}(x) = O(x^{-(2|F|-2)\sigma-\epsilon})$ as $x \to \infty$ for some $\epsilon > 0$. In this case, we have

$$\frac{N_\alpha^{(F)} - E(N_\alpha^{(F)})}{\sqrt{\text{var}(N_\alpha^{(F)})}} \to \mathcal{N}(0,1). \tag{41}$$

Moreover, if $\int w^6 \rho(dw) < \infty$*, then Assumptions 4 and 5 also hold. It follows that Theorems 15, 17 and 18 apply and, for any* $\sigma \in [0,1]$ *and any* ℓ*,*

$$\frac{N_\alpha - E(N_\alpha)}{\sqrt{\mathrm{var}(N_\alpha)}} \to \mathcal{N}(0,1). \tag{42}$$

Finally, assume $\sigma \in (0,1)$ *and* $\widetilde{\ell}(t) = c > 0$*. If additionally*

$$\overline{\rho}(x) \sim c_0 x^{-\sigma\tau} \text{ as } x \to \infty \tag{43}$$

for some $\tau > 0, c_0 > 0$ *then Assumption 3 is also satisfied with* $\tau > 0$*,* $\ell_2(x) = \frac{c_0}{2^{\sigma\tau} c^\tau \Gamma(1-\sigma)^\tau}$ *and Proposition 7 applies; that is, for fixed* α

$$E(N_{\alpha,j}) \sim \frac{\alpha^{1+\tau} \tau \ell_2(j)}{j^{1+\tau}} \text{ as } j \to \infty.$$

We consider below two specific choices of mean measures ρ. Both measures have similar properties for large graph size α, but different properties for large degrees j.

Generalised Gamma measure. Let ρ be the generalised gamma measure

$$\rho(dw) = 1/\Gamma(1-\sigma_0) w^{-1-\sigma_0} e^{-\tau_0 w} dw \tag{44}$$

with $\tau_0 > 0$ and $\sigma_0 \in (-\infty, 1)$. The tail Lévy intensity satisfies

$$\overline{\rho}(x) \sim \begin{cases} \frac{1}{\Gamma(1-\sigma_0)\sigma_0} x^{-\sigma_0} & \sigma_0 > 0 \\ \log(1/x) & \sigma_0 = 0 \\ -\frac{\tau_0^{\sigma_0}}{\sigma_0} & \sigma_0 < 0 \end{cases}$$

as $x \to 0$. Then for $\sigma_0 \in (0,1)$ (sparse with power-law)

$$N_\alpha^{(e)} \asymp N_\alpha^{2/(1+\sigma_0)}, \quad \frac{N_{\alpha,j}}{N_\alpha} \to \frac{\sigma_0 \Gamma(j-\sigma_0)}{j! \Gamma(1-\sigma_0)}, \quad j \geq 1.$$

For $\sigma_0 = 0$ (sparse, almost dense), $N_\alpha^{(e)} \asymp N_\alpha^2 / \log(N_\alpha)^2$ and $N_{\alpha,j}/N_\alpha \to 0, j \geq 1$; for $\sigma_0 < 0$ (dense) $N_\alpha^{(e)} \asymp N_\alpha^2$ and $N_{\alpha,j}/N_\alpha \to 0, \quad j \geq 1$ almost surely as α tends to infinity. The constants in the asymptotic results are omitted for simplicity of exposure but can be obtained as well from the results of Section 3. $\int w^p \rho(dw) < \infty$ for all $p \geq 1$, hence the global clustering coefficient converges, and the CLT applies for the number of subgraphs and the number of nodes. Note that Equation (43) is not satisfied, as the Lévy measure has exponentially decaying tails, and Proposition 7 does not apply. The asymptotic properties of this model are illustrated in Figure 2 for $\sigma_0 = 0.2$ and $\tau_0 = 2$ (sparse, power-law regime).

Generalised gamma Pareto measure. Consider the generalised gamma Pareto measure, introduced by Ayed et al. (2019, 2020)

$$\rho(dw) = \frac{1}{\Gamma(1-\sigma)} w^{-1-\sigma\tau} \gamma(\sigma(\tau-1), \beta w) dw$$

where $\gamma(s,x) = \int_0^x u^{s-1}e^{-u}du$ is the lower incomplete gamma function, $c > 0$, $\tau > 1$, $\sigma \in (0,1)$. The tail Lévy intensity satisfies

$$\overline{\rho}(x) \sim cx^{-\sigma} \text{ as } x \to 0$$
$$\overline{\rho}(x) \sim c_0 x^{-\sigma\tau} \text{ as } x \to \infty$$

where $c = \frac{\beta^{\sigma(\tau-1)}}{\sigma^2(\tau-1)\Gamma(1-\sigma)}$ and $c_0 = \frac{\Gamma(\sigma(\tau-1))}{\sigma\tau\Gamma(1-\sigma)}$. It is both regularly varying at 0 and infinity and satisfies (40) and (43). We therefore have, almost surely,

$$N_\alpha^{(e)} \asymp N_\alpha^{2/(1+\sigma)}, \quad \frac{N_{\alpha,j}}{N_\alpha} \to \frac{\sigma_0\Gamma(j-\sigma)}{j!\Gamma(1-\sigma)}, \quad j \geq 1.$$

Proposition 7 applies and, for large degree nodes,

$$E(N_{\alpha,j}) \sim \frac{\tau\alpha^{1+\tau}c_0}{2^{\sigma\tau}c^\tau\Gamma(1-\sigma)^\tau} \frac{1}{j^{1+\tau}} \text{ as } j \to \infty.$$

The global clustering coefficient converges if $\tau > 2$, and the CLT applies for the number of subgraphs F if $\tau > 2|F| - 2$, and for the number of nodes if $\sigma\tau > 6$.

6.6. Proof of Proposition 23

The marginal graphon function is given by $\mu(x) = \psi(2\overline{\rho}^{-1}(x))$ where $\psi(t) = \int_0^\infty (1-e^{-wt})\rho(dw)$ is the Laplace exponent. Its generalised inverse is given by $\mu^{-1}(x) = \overline{\rho}(\psi^{-1}(x)/2)$. The Laplace exponent satisfies $\psi(t) \sim mt$ as $t \to 0$. It therefore follows that μ^{-1} satisfies Assumption 1 with $\ell(x) = (2m)^\sigma \widetilde{\ell}(x)$. Ignoring loops, the model is of the form given by Equation (18) with $f(x) = 2m\overline{\rho}^{-1}(x)$. Assumption 2 is therefore satisfied. Regarding the global clustering coefficient, $\int \psi(2w)^2\rho(dw) \leq 4\int w^2\rho(dw) < \infty$ so its limit is finite. For the local clustering coefficient, using dominated convergence and the inequality $\frac{1-e^{-2\overline{\rho}^{-1}(x)y}}{2\overline{\rho}^{-1}(x)} \leq y$, we obtain

$$\begin{aligned}\int W(x,y)W(y,z)W(x,z)dydz &= \int (1-e^{-2\overline{\rho}^{-1}(x)y})(1-e^{-2\overline{\rho}^{-1}(x)z})(1-e^{-2yz})\rho(dy)\rho(dz) \\ &\sim 4\overline{\rho}^{-1}(x)^2 \int yz(1-e^{-2yz})\rho(dy)\rho(dz)\end{aligned}$$

Using the fact that $\mu(x) = \psi(2\overline{\rho}^{-1}(x)) \sim 2m\overline{\rho}^{-1}(x)$ as $x \to \infty$, we obtain the result. Finally, if $\overline{\rho}$ satisfies (40), then $\psi(t) \sim \Gamma(1-\sigma)\widetilde{\ell}(t)t^\sigma$ as $t \to \infty$. Using (Bingham et al., 1987, Proposition 1.5.15)

$$\psi^{-1}(t) \sim \Gamma(1-\sigma)^{-1/\sigma}\widetilde{\ell}^{\#1/\sigma}(t^{1/\sigma})t^{1/\sigma}$$

as $t \to \infty$, where $\widetilde{\ell}^{\#}$ is the de Bruijn conjugate of $\widetilde{\ell}$. We obtain $\psi^{-1}(t) = \ell_3(t^{1/\sigma})t^{\frac{1}{\sigma}}$ where ℓ_3 is a slowly varying function with$\ell_3(t^{1/\sigma}) \sim \widetilde{\ell}^{\#1/\sigma}(t^{1/\sigma})\Gamma(1-\sigma)^{-1/\sigma}$ as $t \to \infty$. We therefore have$\mu^{-1}(t) \sim c_0 2^{-\tau\sigma}\ell_3(t^{1/\sigma})^{\sigma\tau}t^{\tau}$ as $t \to \infty$. If $\widetilde{\ell}(t) = c$, then $\ell_3(t) = (c\Gamma(1-\sigma))^{-1/\sigma}$.

For the CLT for the number of subgraphs F to hold, we need $\int_0^\infty \mu(x)^{2|F|-2}dx < \infty$. As μ is monotone decreasing and integrable, we only need $\mu(x)^{2|F|-2} = \psi(2\overline{\rho}^{-1}(x))^{2|F|-2}$ to be integrable in a neighbourhood of 0. In the dense case, $\psi(t)$ is bounded, and the condition holds. If $\overline{\rho}$ satisfies (40), then $\psi(t) \sim \Gamma(1-\sigma)\widetilde{\ell}(t)t^{\sigma}$ as $t \to \infty$. For $\sigma \in (0,1]$ (sparse regime) the condition holds if $\overline{\rho}(x) = O(x^{-(2|F|-2)\sigma-\epsilon})$ as $x \to \infty$ for some $\epsilon > 0$.

We now check the assumptions for the CLT for the number of nodes. Noting again that $\mu(x) \sim 2m\overline{\rho}^{-1}(x)$ as $x \to \infty$, we have, using the inequality $1 - e^{-x} \leq x$,

$$\begin{aligned}\nu(x,y) &= \int \left(1 - e^{-2\overline{\rho}^{-1}(x)w}\right)\left(1 - e^{-2\overline{\rho}^{-1}(y)w}\right)\rho(dw) \\ &\leq L(x)L(y)\mu(x)\mu(y)\end{aligned}$$

where $L(x) = 2\frac{\overline{\rho}^{-1}(x)}{\mu(x)}\sqrt{\int w^2\rho(dw)} \to \sqrt{\int w^2\rho(dw)}/m$ as $x \to \infty$. Using now the inequality $1 - e^{-x} \geq xe^{-x}$, we have

$$\nu(x,y) \geq 4\overline{\rho}^{-1}(x)\overline{\rho}^{-1}(y)\int w^2 e^{-2(\overline{\rho}^{-1}(x)+\overline{\rho}^{-1}(y))w}\rho(dw)$$

As $\int w^2 e^{-2(\overline{\rho}^{-1}(x)+\overline{\rho}^{-1}(y))w}\rho(dw) \to \int w^2\rho(dw)$ as $\min(x,y) \to \infty$, there is $C_0 = 2\int w^2\rho(dw)$ and x_0 such that for all $x, y > x_0$, $\nu(x,y) \geq C_0\mu(x)\mu(y)$.

More generally, if $\int w^6\rho(dw) < \infty$, then for any $j \leq 6$

$$\int_0^\infty \prod_{i=1}^{j} W(x_i, y)dy \leq \prod_{i=1}^{j} L(x_i)\mu(x_i)$$

where $L(x) = 2\frac{\overline{\rho}^{-1}(x)}{\mu(x)}\max\left(1, \max_{j=1,\dots,6}\int w^j\rho(dw)\right) \to \max\left(1, \max_{j=1,\dots,6}\int w^j\rho(dw)\right)/m$ as $x \to \infty$. Note also that $\int L(x)\mu(x)dx = 2\max\left(1, \max_{j=1,\dots,6}\int w^j\rho(dw)\right)\int w\rho(dw) < \infty$.

7. Sparse and dense models with local structure

In this section, we develop a class of models which allows to control separately the local structure, for example the presence of communities or particular subgraphs, and the global sparsity/power-law properties. The class of models introduced can be used as a way of sparsifying any dense graphon model.

7.1. Statement of the results

Due to Kallenberg's representation theorem, any exchangeable point process can be represented by Equation (2). However, it may be more suitable to use a different formulation where the

function W is defined on a general space, not necessarily $\mathbb{R}_+^2$, as discussed by Borgs et al. (2018). Such a construction may lead to more interpretable parameters and easier inference methods. Indeed, a few sparse vertex-exchangeable models, such as the models of Herlau et al. (2016) or Todeschini et al. (2020) are written in a way such that it is not straightforward to express them in the form given by (2).

In this section we show that the above results easily extend to models expressed in the following way. Let F be a probability space. Writing $\vartheta = (u, v) \in \mathbb{R}_+ \times F$, let $\xi(d\vartheta) = duG(dv)$ where G is some probability distribution on F. Consider models expressed as in (1) with

$$Z_{ij} \mid (\theta_k, \vartheta_k)_{k=1,2,\ldots} \sim \text{Bernoulli}\{W\ (\vartheta_i, \vartheta_j)\}, \quad W : (\mathbb{R}_+ \times F)^2 \to [0,1] \tag{45}$$

where $(\theta_k, \vartheta_k)_{k=1,2,\infty}$ are the points of a Poisson point process with mean measure $d\theta\xi(d\vartheta)$ on $\mathbb{R}_+ \times (\mathbb{R}_+ \times F)$. Let us assume additionally that the function W factorizes in the following way

$$W((u_i, v_i), (u_j, v_j)) = \omega(v_i, v_j)\eta(u_i, u_j). \tag{46}$$

where $\omega : F \times F \to [0,1]$ and the function $\eta : \mathbb{R}_+ \times \mathbb{R}_+ \to [0,1]$ is integrable. In this model ω can capture the local structure, as in the classical dense graphon, and η the sparsity behaviour of the graph. Let $\mu_\eta(u) = \int_0^\infty \eta(u, u')du'$, $\mu_\omega(v) = \int_F \omega(v, v')G(dv')$ and $\nu_\eta(x, y) = \int_{\mathbb{R}_+^2} \eta(x, z)\eta(y, z)dz$. The results presented in Section 3 remain valid when μ_η and ν_η satisfy Assumptions 1 and 2. The proof of Proposition 24 is given in Section 7.2.

Proposition 24 *Consider the model defined by Equations (45) and (46) and assume that the functions μ_η and ν_η satisfy assumptions 1 and 2. Then the conclusions of Proposition 2 hold and so do the conclusions of Theorems 3 and 4 with $\ell(\alpha)$ and $\ell_1(\alpha)$ replaced respectively by*

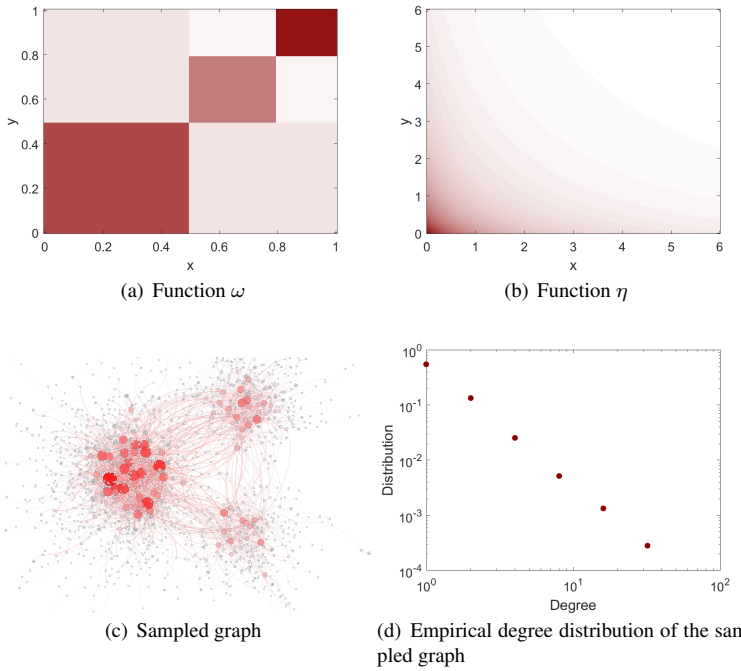
$$\tilde{\ell}(\alpha) = \ell(\alpha) \int_F \mu_\omega(v)^\sigma G(dv), \quad \tilde{\ell}_1(\alpha) = \ell_1(\alpha) \int_F \mu_\omega(v)^\sigma G(dv).$$

Consider for example the following class of models for sparse and dense stochastic block-models.

Example 25 (Dense and Sparse stochastic block-models) *Consider $F = [0,1]$ and G the uniform distribution on $[0,1]$. We choose for ω the graphon function associated to a (dense) stochastic block-model. For some partition $A_1, \ldots, A_p$ of $[0,1]$, and any $v, v' \in [0,1]$, let*

$$\omega(v, v') = B_{k,\ell} \tag{47}$$

with $v \in A_k$, $v' \in A_\ell$ and B is a $p \times p$ matrix where $B_{k,\ell} \in [0,1]$ denotes the probability that a node in community k forms a link with a node in community ℓ. ω defines the community structure of the graph, and η will tune its sparsity properties. Choosing $\eta(x, y) = \mathbb{1}_{x \le 1}\mathbb{1}_{y \le 1}$ yields the dense, standard stochastic block-model. Choosing $\eta(x, y) = \exp(-x - y)$ yields a sparse stochastic block-model without power-law behaviour, etc. An illustration of this model to obtain sparse stochastic block-models with power-law behaviour, generalizing the model of Section 6.3, is given in Figure 4. The function ω is defined by: $A_1 = [0, 0.5)$, $A_2 = [0.5, 0.8)$, $A_3 = [0.8, 1]$, $B_{11} = 0.7, B_{22} = 0.5, B_{33} = 0.9, B_{12} = B_{13} = 0.1, B_{23} = 0.05$ and $\eta(x, y) = (1 + x)^{-1/\sigma}(1 + y)^{-1/\sigma}$, with $\sigma = 0.8$.



(a) Function ω (b) Function η

(c) Sampled graph (d) Empirical degree distribution of the sampled graph

Figure 4. Illustration of a sparse stochastic block-model with 3 communities. (a) Function ω, that controls the local community structure. A darker color represents a higher value. (b) Function η, that controls the sparsity. (c) Graph sampled from the sparse stochastic block-model using $\alpha = 50$. The size of each node are proportional to its degree. (d) Empirical degree distribution of the sampled graph.

More generally, one can build on the large literature on (dense) graphon/exchangeable graph models, and combine these models with a function η satisfying Assumptions 1 and 2, such as those described in the previous section, in order to sparsify a dense graphon and control its sparsity/power-law properties.

Remark 26 *We can also obtain asymptotic results for those functions W that do not satisfy the separability condition (46). Let $\mu(u,v) = \int_{\mathbb{R}_+ \times F} W((u,v),(u',v'))du'dv'$. Assume that, for each fixed v, there exists $u_0(v) > 0$ such that for $u > u_0$*

$$C_3 \tilde{\mu}_\eta(u)\tilde{\mu}_\omega(v) \leq \mu(u,v) \leq C_4 \tilde{\mu}_\eta(u)\tilde{\mu}_\omega(v) \tag{48}$$

where $\tilde{\mu}_\omega : F \to \mathbb{R}_+$, $\tilde{\mu}_\eta : \mathbb{R}_+ \to \mathbb{R}_+$ with $\tilde{\mu}_\eta(u) = \int_0^\infty \tilde{\eta}(u,u')du'$ for some positive function $\tilde{\eta}$, and $C_3 > 0$ and $C_4 > 0$. Assume that $\tilde{\mu}_\eta$ and $\tilde{\nu}_\eta$ verify Assumptions 1 and 2. Then the results of Theorems 3 and 4, Corollary 5 hold up to a constant. For example, we have for $\sigma \in [0,1]$, $N_\alpha^{(e)} \asymp N_\alpha^{2/(1+\sigma)} \ell_\sigma^*(N_\alpha)$ almost surely as α tends to infinity. In particular, the inequality from

(48) *is satisfied if*

$$W((u_i, v_i), (u_j, v_j)) = 1 - e^{-\tilde{\omega}(v_i, v_j)\bar{\eta}(u_i, u_j)}. \tag{49}$$

The models developed by Herlau et al. (2016) and Todeschini et al. (2020) for capturing (overlapping) communities fit in this framework. Ignoring loops, both models can be written under the form given by Equation (49) with $\tilde{\eta}(u, u') = 2\overline{\rho}^{-1}(u)\overline{\rho}^{-1}(u')$*, where* ρ *is a Lévy measure on* $(0, +\infty)$ *and* $\overline{\rho}(x) = \int_x^\infty \rho(dw)$ *is the tail Lévy intensity with generalised inverse* $\overline{\rho}^{-1}(x)$*. When* $\tilde{\omega}$ *is given by Equation (47), it corresponds to the (dense) stochastic blockmodel graphon of Herlau et al. (2016) and if* $\tilde{\omega}(v_i, v_j) = v_i^T v_j$ *with* $v_i \in \mathbb{R}_+^p$*, it corresponds to the model of Todeschini et al. (2020). For instance, let* ρ *be the mean measure from Equation (44) with parameters* $\tau_0 > 0$ *and* $\sigma_0 \in (-\infty, 1)$*. Then for* $\sigma_0 \in (0, 1)$*, the corresponding sparse regime with power-law for this graph is given by*

$$N_\alpha^{(e)} \asymp N_\alpha^{2/(1+\sigma_0)}, \quad \frac{C_3}{C_4}\frac{\sigma_0\Gamma(j-\sigma_0)}{j!\Gamma(1-\sigma_0)} \leq \lim_{\alpha\to\infty}\frac{N_{\alpha,j}}{N_\alpha} \leq \frac{C_4}{C_3}\frac{\sigma_0\Gamma(j-\sigma_0)}{j!\Gamma(1-\sigma_0)}, \quad j \geq 1$$

For $\sigma_0 = 0$ *(sparse, almost dense regime)* $N_\alpha^{(e)} \asymp N_\alpha^2/\log(N_\alpha)^2$ *and* $N_{\alpha,j}/N_\alpha \to 0, j \geq 1$*; for* $\sigma_0 < 0$ *(dense regime)* $N_\alpha^{(e)} \asymp N_\alpha^2$ *and* $N_{\alpha,j}/N_\alpha \to 0, j \geq 1$ *almost surely as* α *tends to infinity.*

7.2. Proof of Proposition 24

The proofs of Proposition 2 and Theorems 3 and 4 hold with x replaced by $(u, v) \in \mathbb{R}_+ \times F$, $dx = duG(dv)$ and $\mu(x) = \mu_\eta(u)\mu_\omega(v)$. We thus need only prove that if η verifies Assumptions 1 and 2 then Lemmas B.2, B.3 and B.4 in Appendix hold. Recall that $\mu(x) = \mu_\eta(u)\mu_\omega(v)$, for $x = (u, v)$. Then for all v such that $\mu_\omega(v) > 0$ we apply Lemma B.2 to

$$g_0(t) = \int_0^\infty (1 - e^{-t\mu_\eta(u)})du, \quad g_r(t) = \int_0^\infty \mu_\eta(u)^r e^{-t\mu_\eta(u)}du, \quad t = \alpha\mu_\omega(v).$$

This leads to, for all v such that $\mu_\omega(v) > 0$

$$\begin{aligned}\int_0^\infty (1 - e^{-\alpha\mu_\omega(v)\mu_\eta(u)})du &= \Gamma(1-\sigma)\alpha^\sigma\ell(\alpha)\mu_\omega(v)^\sigma\frac{\ell\{\alpha\mu_\omega(v)\}}{\ell(\alpha)}\{1 + o(1)\} \\ &= \Gamma(1-\sigma)\alpha^\sigma\ell(\alpha)\mu_\omega(v)^\sigma\{1 + o(1)\}.\end{aligned}$$

To prove that there is convergence in $L_1(G)$, note that if $\mu_\omega(v) > 0$ and since $\mu_\omega \leq 1$,

$$\int_0^\infty (1 - e^{-\alpha\mu_\omega(v)\mu_\eta(u)})du = \int_0^\infty \mu_\eta^{-1}\left\{\frac{z}{\alpha\mu_\omega(v)}\right\}e^{-z}dz \leq \int_0^\infty \mu_\eta^{-1}\left(\frac{z}{\alpha}\right)e^{-z}dz.$$

Moreover

$$\sup_{\alpha\geq 1}\frac{1}{\alpha^\sigma\ell(\alpha)}\int_0^\infty \mu_\eta^{-1}\left(\frac{z}{\alpha}\right)e^{-z}dz < +\infty,$$

thus the Lebesgue dominated convergence theorem implies

$$\int_F\int_0^\infty (1 - e^{-\alpha\mu_\omega(v)\mu_\eta(u)})duG(dv) \sim \Gamma(1-\sigma)\alpha^\sigma\ell(\alpha)\int_F \mu_\omega(v)^\sigma G(dv)$$

when $\sigma < 1$ and when $\sigma = 1$,

$$\int_F \int_0^\infty (1 - e^{-\alpha\mu_\omega(v)\mu_\eta(u)}) du G(dv) \sim \alpha \ell_1(\alpha) \int_F \mu_\omega(v) G(dv).$$

The same reasoning is applied to the integrals

$$\int_F \mu_\omega(v)^r \int_0^\infty \mu_\eta(u)^r e^{-\alpha\mu_\omega(v)\mu_\eta(u)}) du G(dv).$$

To verify Lemma B.3, note that

$$h_0(\alpha) = \int_F \omega(v,v) \int_0^\infty \eta(u,u)(1 - e^{-\alpha\mu_\omega(v)\mu_\eta(u)}) du G(dv),$$
$$h_r(\alpha) = \int_F \omega(v,v)\mu_\omega(v)^r \int_0^\infty \eta(u,u)\mu_\eta(u)^r e^{-\alpha\mu_\omega(v)\mu_\eta(u)} du G(dv)$$

so that the Lebesgue dominated convergence Theorem also leads to

$$h_0(\alpha) \sim \int_F \omega(v,v) \int_0^\infty \eta(u,u) du G(dv), \quad h_r(\alpha) = o(\alpha^{-r})$$

and the control of the integrals $\int_{\mathbb{R}_+ \times F} \{t\mu(u,v)\} e^{-t\mu(u,v)} du G(dv)$ as in Lemma B.4.

8. Conclusion

In this article, we derived a number of properties of graphs based on exchangeable random measures. We relate the sparsity and power-law properties of the graphs to the regular variation properties of the marginal graphon function, identifying four different regimes, from dense to almost extremely sparse. We derived asymptotic results for the global and local clustering coefficients. We derived a central limit theorem for the number of nodes N_α in the sparse and dense regimes, and for the number of nodes of degree greater than j in the dense regime. We conjecture that a CLT also holds for $N_{\alpha,j}$ in the sparse regime, under assumptions similar to Assumptions 4 and 5, and that a (lengthy) proof similar to that of Theorem 18 could be used. We leave this for future work.

Acknowledgment

The authors thank Zacharie Naulet for helpful feedback and suggestions on an earlier version of this article. The project leading to this work has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 834175).

References

Aldous, D. J. (1981). Representations for partially exchangeable arrays of random variables. *Journal of Multivariate Analysis 11*(4), 581–598.

Ayed, F., J. Lee, and F. Caron (2019). Beyond the Chinese restaurant and Pitman-Yor processes: Statistical models with double power-law behavior. In *International Conference on Machine Learning*, pp. 395–404.

Ayed, F., J. Lee, and F. Caron (2020). The normal-generalised gamma-pareto process: A novel pure-jump l\'evy process with flexible tail and jump-activity properties. *arXiv preprint arXiv:2006.10968*.

Bickel, P. J. and A. Chen (2009). A nonparametric view of network models and Newman–Girvan and other modularities. *Proceedings of the National Academy of Sciences 106*(50), 21068–21073.

Bickel, P. J., A. Chen, and E. Levina (2011). The method of moments and degree distributions for network models. *The Annals of Statistics 39*(5), 2280–2301.

Bingham, N. H., C. M. Goldie, and J. L. Teugels (1987). *Regular variation*, Volume 27. Cambridge university press.

Bollobás, B., S. Janson, and O. Riordan (2007). The phase transition in inhomogeneous random graphs. *Random Structures & Algorithms 31*(1), 3–122.

Bollobás, B. and O. Riordan (2009). Metrics for sparse graphs. In S. Huczynska, J. Mitchell, and C. Roney-Dougal (Eds.), *Surveys in combinatorics*. arXiv:0708.1919: Cambridge University Press.

Borgs, C., J. T. Chayes, H. Cohn, and N. Holden (2018). Sparse exchangeable graphs and their limits via graphon processes. *Journal of Machine Learning Research 18*, 1–71.

Borgs, C., J. T. Chayes, H. Cohn, and V. Veitch (2019). Sampling perspectives on sparse exchangeable graphs. *The Annals of Probability 47*(5), 2754–2800.

Borgs, C., J. T. Chayes, S. Dhara, and S. Sen (2019). Limits of sparse configuration models and beyond: Graphexes and multi-graphexes. *arXiv preprint arXiv:1907.01605*.

Caron, F. and E. Fox (2017). Sparse graphs using exchangeable random measures. *Journal of the Royal Statistical Society B 79*, 1–44. Part 5.

Caron, F., F. Panero, and J. Rousseau (2020). On sparsity, power-law and clustering properties of graphs based of graphex processes:supplementary material. Technical report, University of Oxford.

Chatterjee, S. (2015). Matrix estimation by universal singular value thresholding. *The Annals of Statistics 43*(1), 177–214.

Diaconis, P. and S. Janson (2008). Graph limits and exchangeable random graphs. *Rendiconti di Matematica e delle sue Applicazioni. Serie VII*, 33–61.

Gao, C., Y. Lu, and H. Zhou (2015). Rate-optimal graphon estimation. *The Annals of Statistics 43*(6), 2624–2652.

Gnedin, A., B. Hansen, and J. Pitman (2007). Notes on the occupancy problem with infinitely many boxes: general asymptotics and power laws. *Probability Surveys 4*(146-171), 88.

Herlau, T., M. N. Schmidt, and M. Mørup (2016). Completely random measures for modelling block-structured sparse networks. In *Advances in Neural Information Processing Systems 29 (NIPS 2016)*.

Hoover, D. N. (1979). Relations on probability spaces and arrays of random variables. *Preprint, Institute for Advanced Study, Princeton, NJ*.

Janson, S. (2016). Graphons and cut metric on sigma-finite measure spaces. *arXiv:1608.01833*.

Janson, S. (2017). On convergence for graphexes. *arXiv preprint arXiv:1702.06389*.

Kallenberg, O. (1990). Exchangeable random measures in the plane. *Journal of Theoretical Probability 3*(1), 81–136.

Kolaczyk, E. D. (2009). *Statistical Analysis of Network Data. Methods and models.* Springer.

Last, G., G. Peccati, and M. Schulte (2016). Normal approximation on Poisson spaces: Mehler's formula, second order Poincaré inequalities and stabilization. *Probability theory and related fields 165*(3-4), 667–723.

Latouche, P. and S. Robin (2016). Variational Bayes model averaging for graphon functions and motif frequencies inference in W-graph models. *Statistics and Computing 26*(6), 1173–1185.

Lloyd, J., P. Orbanz, Z. Ghahramani, and D. Roy (2012). Random function priors for exchangeable arrays with applications to graphs and relational data. In *Advances in Neural Information Processing Systems 25 (NIPS 2012)*.

Loève, M. (1977). *Probability Theory I (4th ed.)* (4th ed. ed.). New York: Springer-Verlag.

Lovász, L. and B. Szegedy (2006). Limits of dense graph sequences. *Journal of Combinatorial Theory, Series B 96*(6), 933–957.

Naulet, Z., E. Sharma, V. Veitch, and D. M. Roy (2017). An estimator for the tail-index of graphex processes. *arXiv preprint arXiv:1712.01745*.

Newman, M. E. J. (2010). *Networks: An Introduction*. Oxford University Press.

Nowicki, K. and T. Snijders (2001). Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association 96*(455), 1077–1087.

Orbanz, P. and D. M. Roy (2015). Bayesian models of graphs, arrays and other exchangeable random structures. *IEEE Transactions on Pattern Analysis and Machine Intelligence 37*(2), 437–461.

Palla, G., L. Lovász, and T. Vicsek (2010). Multifractal network generator. *Proceedings of the National Academy of Sciences 107*(17), 7640–7645.

Penrose, M. (2003). *Random geometric graphs*, Volume 5. Oxford University Press.

Reitzner, M. and M. Schulte (2013, 11). Central limit theorems for U-statistics of Poisson point processes. *Ann. Probab. 41*(6), 3879–3909.

Resnick, S. (1987). *Extreme values, point processes and regular variation*. Springer-Verlag, New York.

Todeschini, A., X. Miscouridou, and F. Caron (2020). Exchangeable random measures for sparse and modular graphs with overlapping communities. *Journal of the Royal Statistical Society series B*. to appear.

Veitch, V. and D. M. Roy (2015). The class of random graphs arising from exchangeable random measures. *arXiv:1512.03099*.

Veitch, V. and D. M. Roy (2019). Sampling and estimation for (sparse) exchangeable graphs. *Annals of Statistics 47*(6), 3274–3299.

Willmot, G. E. (1990). Asymptotic tail behaviour of poisson mixtures by applications. *Advances in Applied Probability 22*(1), 147–159.

Wolfe, P. J. and S. C. Olhede (2013). Nonparametric graphon estimation. *ArXiv preprint arXiv:1309.5936*.

A. Proofs of Theorem 3 and Proposition 11

Let $g_{\alpha,x}(\theta,\vartheta)$ be defined, for any $\alpha, x, \theta, \vartheta > 0$, by

$$g_{\alpha,x}(\theta,\vartheta) = -\log\{1 - W(x,\vartheta)\}\mathbb{1}_{\theta\leq\alpha}. \tag{50}$$

A.1. Proof of Theorem 3

The mean number of nodes is (Veitch and Roy, 2015, Theorem 5.4)

$$E(N_\alpha) = \alpha\int_{\mathbb{R}_+}\{1 - e^{-\alpha\mu(x)}\}dx + \alpha\int_{\mathbb{R}_+} W(x,x)e^{-\alpha\mu(x)}dx.$$

By the Lebesgue dominated convergence, we have $\alpha\int_{\mathbb{R}_+} W(x,x)e^{-\alpha\mu(x)}dx = o(\alpha)$. We have, using Lemma B.2, for $\sigma \in [0,1)$, as α goes to infinity $\int_{\mathbb{R}_+}(1 - e^{-\alpha\mu(x)})dx \sim \alpha^\sigma\ell(\alpha)\Gamma(1-\sigma)$, and for $\sigma = 1$, $\int_{\mathbb{R}_+}\{1 - e^{-\alpha\mu(x)}\}dx \sim \alpha\ell_1(\alpha)$. It follows that, as α goes to infinity

$$E(N_\alpha) \sim \begin{cases} \alpha^{\sigma+1}\ell(\alpha)\Gamma(1-\sigma) & \text{if } \sigma \in [0,1) \\ \alpha^2\ell_1(\alpha) & \text{if } \sigma = 1 \end{cases}.$$

The mean number of nodes of degree j is (Veitch and Roy, 2015, Theorem 5.5)

$$E(N_{\alpha,j}) = \frac{\alpha^{j+1}}{j!}\int_{\mathbb{R}_+}(1 - W(\vartheta,\vartheta))e^{-\alpha\mu(\vartheta)}\mu(\vartheta)^j d\vartheta + \frac{\alpha^j}{j-1!}\int_{\mathbb{R}_+} e^{-\alpha\mu(\vartheta)}W(\vartheta,\vartheta)\mu(\vartheta)^{j-1}d\vartheta \tag{51}$$

Lemma B.3, implies that

$$-\frac{\alpha^{j+1}}{j!}\int_{\mathbb{R}_+} W(\vartheta,\vartheta)e^{-\alpha\mu(\vartheta)}\mu(\vartheta)^j d\vartheta + \frac{\alpha^j}{j-1!}\int_{\mathbb{R}_+} e^{-\alpha\mu(\vartheta)}W(\vartheta,\vartheta)\mu(\vartheta)^{j-1}d\vartheta = o(\alpha)$$

and from Lemma B.2, we have, when $\sigma \in [0,1)$

$$\frac{\alpha^{j+1}}{j!}\int_{\mathbb{R}_+} e^{-\alpha\mu(\vartheta)}\mu(\vartheta)^j d\vartheta \sim \frac{\sigma\Gamma(j-\sigma)}{j!}\alpha^{1+\sigma}\ell(\alpha).$$

If $\sigma = 1$, then $\alpha^2\int_{\mathbb{R}_+} e^{-\alpha\mu(\vartheta)}\mu(\vartheta)d\vartheta \sim \alpha^2\ell_1(\alpha)$ and for $j \geq 2$

$$\frac{\alpha^{j+1}}{j!}\int_{\mathbb{R}_+} e^{-\alpha\mu(\vartheta)}\mu(\vartheta)^j d\vartheta \sim \frac{1}{j(j-1)}\alpha^2\ell(\alpha).$$

We finally obtain, for $\sigma \in [0,1)$ $E(N_{\alpha,j}) \sim \frac{\sigma\Gamma(j-\sigma)}{j!}\alpha^{1+\sigma}\ell(\alpha)$, and for $\sigma = 1$, $E(N_{\alpha,1}) \sim \alpha^2\ell_1(\alpha)$, and $E(N_{\alpha,j}) \sim \alpha^2/\{j(j-1)\}\ell(\alpha)$, for $j \geq 2$.

A.2. Proof of Proposition 11

For $j \geq 1$, define

$$R_{\alpha j} = \sum_i T_{\alpha i} \mathbb{1}_{D_{\alpha i = j}}. \tag{52}$$

$R_{\alpha j}$ corresponds to the number of triangles having a node of degree j as a vertex, where triangles having $k \leq 3$ degree-j nodes as vertices are counted k times. We therefore have

$$C^{(\ell)}_{\alpha,j} = \frac{2}{j(j-1)} \frac{R_{\alpha j}}{N_{\alpha,j}}.$$

The proof for the asymptotic behaviour of the local clustering coefficients $C^{(\ell)}_{\alpha,j}$ is organised as follows. We first derive a convergence result for $E(R_{\alpha j})$. This result is then extended to an almost sure result. The extension requires some additional work as $R_{\alpha j}$ is not monotone, and $\sum_{j \geq k} R_{\alpha k}$ is monotone but not of the same order as $R_{\alpha j}$, hence a proof similar to that for $N_{\alpha j}$ (see Section 3.2) cannot be used. The almost sure convergence results for $C^{(\ell)}_{\alpha,j}$ and $\overline{C}^{(\ell)}_{\alpha}$ then follow from the almost sure convergence result for $R_{\alpha j}$.

We have

$$R_{\alpha j} = \sum_i T_{\alpha i} \mathbb{1}_{D_{\alpha i = j}} = \frac{1}{2} \sum_{i \neq l \neq k} Z_{il} Z_{ik} Z_{lk} \mathbb{1}_{\sum_s Z_{is} = j} \mathbb{1}_{\theta_s \leq \alpha} \mathbb{1}_{\theta_i \leq \alpha} \mathbb{1}_{\theta_l \leq \alpha} \mathbb{1}_{\theta_k \leq \alpha}$$

and

$$\begin{aligned}
E\left(R_{\alpha j} \mid M\right) = \frac{1}{2} \sum_{i \neq l \neq k} & W(\vartheta_i, \vartheta_l) W(\vartheta_i, \vartheta_k) W(\vartheta_l, \vartheta_k) \frac{1}{(j-2)!} \\
& \times \sum_{i_1 \neq i_2 \ldots \neq i_{j-2} \neq l \neq k} \left[\prod_{s=1}^{j-2} W(\vartheta_i, \vartheta_s) \right] e^{-\sum_{s \neq l,k,i_1,\ldots,i_{j-2}} g_{\alpha,\vartheta_i}(\theta_s, \vartheta_s)} \\
= \frac{1}{2\,(j-2)!} & \sum_{i \neq l \neq k \neq i_1 \neq i_2 \ldots \neq i_{j-2}} W(\vartheta_i, \vartheta_l) W(\vartheta_i, \vartheta_k) W(\vartheta_l, \vartheta_k) (1 - W(\vartheta_i, \vartheta_i)) \\
& \times \left[\prod_{s=1}^{j-2} W(\vartheta_i, \vartheta_s) \right] e^{-\sum_{s \neq l,k,i_1,\ldots,i_{j-2}} g_{\alpha,\vartheta_i}(\theta_s, \vartheta_s)} \\
+ \frac{1}{2(j-3)!} & \sum_{i \neq l \neq k \neq i_1 \neq i_2 \ldots \neq i_{j-3}} W(\vartheta_i, \vartheta_i) W(\vartheta_i, \vartheta_l) W(\vartheta_i, \vartheta_k) W(\vartheta_l, \vartheta_k) \\
& \times \prod_{s=1}^{j-3} W(\vartheta_i, \vartheta_s) e^{-\sum_{s \neq i,l,k,i_1,\ldots,i_{j-3}} g_{\alpha,\vartheta_i}(\theta_s, \vartheta_s)}
\end{aligned}$$

where $g_{\alpha,x}(\theta,\vartheta)$ is defined in Equation (50). Applying the Slivnyak-Mecke theorem, we obtain

$$E\left(R_{\alpha j}\right) = \frac{\alpha^{j+1}}{2(j-2)!}\int_{\mathbb{R}_+^3} W(x,y)W(x,z)W(y,z)(1-W(x,x))\mu(x)^{j-2}e^{-\alpha\mu(x)}dxdydz + \frac{\alpha^{j}}{2(j-3)!}\int_{\mathbb{R}_+^3} W(x,y)W(x,z)W(y,z)W(x,x)\mu(x)^{j-3}e^{-\alpha\mu(x)}dxdydz. \quad (53)$$

Note that under Assumption 1 with $\sigma \in (0,1)$, $\mu(x) > 0$ for all x. The leading term in the right-handside of Equation (53) is the first term. We have therefore

$$E\left(R_{\alpha j}\right) \sim \frac{\alpha^{j+1}}{2(j-2)!}\int_{\mathbb{R}_+^3} L(x)\mu(x)^{j}e^{-\alpha\mu(x)}dxdydz$$

where

$$L(x) = \frac{(1-W(x,x))\int_{\mathbb{R}_+^2} W(x,y)W(x,z)W(y,z)dydz}{\mu(x)^2}.$$

As $\lim_{x\to\infty} W(x,x) = 0$, the condition (28) implies $\lim_{x\to\infty} L(x) = b$.

Case $b > 0$. Assume first that $b > 0$. In this case, L is a slowly varying function by assumption. Therefore, using Lemma B.5, we have, under Assumption 1, for $\sigma \in (0,1)$

$$\int_0^\infty L(x)\mu(x)^{j}e^{-\alpha\mu(x)}dx \sim \sigma b\ell(\alpha)\Gamma(j-\sigma)\alpha^{\sigma-j}.$$

as α tends to infinity. Hence

$$E\left(R_{\alpha j}\right) \sim \frac{b\sigma\Gamma(j-\sigma)}{2(j-2)!}\alpha^{1+\sigma}\ell(\alpha) \quad (54)$$

as α tends to infinity. In order to obtain a convergence in probability, we state the following proposition, whose proof is given in Section S1.3 in the Supplementary Material (Caron et al., 2020) and is similar to that of Proposition 9.

Proposition A.1 *Under Assumptions 1 and 2, with $\sigma \in [0,1]$, slowly varying function ℓ and positive scalar a satisfying (17), we have*

$$\mathrm{var}\left(\sum_i T_{\alpha i}\mathbb{1}_{D_{\alpha i=j}}\right) = O\{\alpha^{3+2\sigma-2a}\ell_\sigma(\alpha)^2\} \text{ as } \alpha\to\infty,$$

and for any sequence α_n going to infinity such that $\alpha_{n+1}-\alpha_n = o(\alpha_n)$,

$$\mathrm{var}\left(\sum_i T_{\alpha_{n+1}i}\mathbb{1}_{D_{\alpha_n i=j}}\mathbb{1}_{\sum_{i'}\mathbb{1}_{\alpha_n<\theta_{i'}\leq\alpha_{n+1}}Z_{ii'}=1}\right) = O\left(\alpha_n^{3+2\sigma-2a}\ell_\sigma(\alpha_n)^2\right) \text{ as } n\to\infty.$$

We now want to find a subsequence α_n along which the convergence is almost sure. Using Chebyshev's inequality and the first part of Proposition A.1, there exists $n_0 \geq 0$ and $C \geq 0$ such that for all $n > n_0$

$$\mathrm{pr}\left(\left|\frac{R_{\alpha_n j}}{E(R_{\alpha_n j})} - 1\right| > \epsilon\right) \leq \frac{C\alpha_n^{3+2\sigma-2a}\ell_\sigma(\alpha_n)^2}{\epsilon^2(\frac{b\sigma\Gamma(j-\sigma)}{2(j-2)!}\alpha_n^{1+\sigma}\ell(\alpha_n))^2}.$$

Now, if Assumption 2 is satisfied for a given $a > 1/2$, consider the sequence

$$\alpha_n = (n\log^2 n)^{1/(2a-1)} \tag{55}$$

so that $\sum_n \alpha_n^{1-2a} < +\infty$ and

$$\sum_n \mathrm{pr}\left(\left|\frac{R_{\alpha_n j}}{E(R_{\alpha_n j})} - 1\right| > \epsilon\right) < \infty.$$

Therefore, using Borel-Cantelli's lemma we have

$$R_{\alpha_n j} \sim \frac{b\sigma\Gamma(j-\sigma)}{2(j-2)!}\alpha_n^{1+\sigma}\ell(\alpha_n)$$

almost surely as $n \to \infty$.

The goal is now to extend this result to $R_{\alpha j}$, by sandwiching. Let $I_\alpha := \{i : \theta_i \leq \alpha\}$. We have the following upper and lower bounds for $R_{\alpha j}$

$$\sum_{i\in I_{\alpha_n}} T_{\alpha_n i}\mathbb{1}_{D_{\alpha i}=j} \leq \sum_{i\in I_\alpha} T_{\alpha i}\mathbb{1}_{D_{\alpha i}=j} \leq \sum_{i\in I_{\alpha_{n+1}}} T_{\alpha_{n+1} i}\mathbb{1}_{D_{\alpha i}=j}. \tag{56}$$

Considering the upper bound of (56):

$$\begin{aligned}\sum_{i\in I_{\alpha_{n+1}}} T_{\alpha_{n+1} i}\mathbb{1}_{D_{\alpha i}=j} &\leq \sum_{i\in I_{\alpha_{n+1}}} T_{\alpha_{n+1} i}\mathbb{1}_{D_{\alpha_{n+1} i}=j} + \sum_{i\in I_{\alpha_{n+1}}} T_{\alpha_{n+1} i}\mathbb{1}_{D_{\alpha i}=j}\mathbb{1}_{D_{\alpha_{n+1} i}>j} \\ &\leq R_{\alpha_{n+1} j} + \widetilde{R}_{nj} \end{aligned} \tag{57}$$

where

$$\widetilde{R}_{nj} = \sum_{i\in I_{\alpha_{n+1}}} T_{\alpha_{n+1} i}\mathbb{1}_{D_{\alpha_n i}\leq j}\mathbb{1}_{\sum_{i'}\mathbb{1}_{\alpha_n<\theta_{i'}\leq\alpha_{n+1}}Z_{ii'}\geq 1}. \tag{58}$$

We can bound the lower bound of (56) by

$$\begin{aligned}\sum_{i\in I_{\alpha_n}} T_{\alpha_n i}\mathbb{1}_{D_{\alpha i}=j} &\geq \sum_{i\in I_{\alpha_n}} T_{\alpha_n i}\mathbb{1}_{D_{\alpha_n i}=j}\mathbb{1}_{D_{\alpha i}=j} \\ &\geq \sum_{i\in I_{\alpha_n}} T_{\alpha_n i}\mathbb{1}_{D_{\alpha_n i}=j} - \sum_{i\in I_{\alpha_n}} T_{\alpha_n i}\mathbb{1}_{D_{\alpha_n i}=j}\mathbb{1}_{D_{\alpha_{n+1} i}>j} \\ &\geq \sum_{i\in I_{\alpha_n}} T_{\alpha_n i}\mathbb{1}_{D_{\alpha_n i}=j} - \sum_{i\in I_{\alpha_{n+1}}} T_{\alpha_{n+1} i}\mathbb{1}_{D_{\alpha_n i}\leq j}\mathbb{1}_{\sum_{i'}\mathbb{1}_{\alpha_n<\theta_{i'}\leq\alpha_{n+1}}Z_{ii'}\geq 1} \\ &= R_{\alpha_n j} - \widetilde{R}_{nj}. \end{aligned} \tag{59}$$

The following Lemma, proved in Section S1.4 of the Supplementary Material (Caron et al., 2020), provides an asymptotic bound for the remainder term $\widetilde{R}_{nj}$.

Lemma A.2 *Let $\widetilde{R}_{nj}$ be defined as in Equation (58). If Assumptions 1 and 2 hold with $\sigma \in (0,1)$ and slowly varying function ℓ, and condition (28) is satisfied with $b > 0$, we have*

$$\widetilde{R}_{nj} = o(\alpha_n^{1+\sigma}\ell(\alpha_n))$$

almost surely as α tends to infinity.

Combining Lemma A.2 with the inequalities (56), (57) and (59), and the fact that $R_{\alpha_n j} \sim R_{\alpha_{n+1} j} \asymp \alpha_n^{1+\sigma}\ell(\alpha_n)$ almost surely as $n \to \infty$, we obtain by sandwiching

$$R_{\alpha j} \sim \frac{b\sigma\Gamma(j-\sigma)}{2(j-2)!}\alpha^{1+\sigma}\ell(\alpha) \text{ almost surely as } \alpha \text{ tends to infinity.}$$

Recalling that $N_{\alpha,j} \sim \frac{\sigma\Gamma(j-\sigma)}{j!}\alpha^{1+\sigma}\ell(\alpha)$ almost surely, we have, for any $j \geq 1$

$$C_{\alpha,j}^{(\ell)} = \frac{2R_{\alpha j}}{j(j-1)N_{\alpha,j}} \to b \text{ almost surely as } \alpha \text{ tends to infinity.}$$

Finally, as $\frac{N_{\alpha,j}}{N_\alpha - N_{\alpha,1}}$ converges to a constant $\pi_j \in (0,1)$ almost surely for any j, we have, using Toeplitz lemma

$$\overline{C}_\alpha^{(\ell)} = \frac{1}{N_\alpha - N_{\alpha,1}}\sum_{j\geq 2} N_{\alpha,j}C_{\alpha,j}^{(\ell)} \to b$$

almost surely as α tends to infinity.

Case $b = 0$. In the case $L(x) \to 0$, Lemma B.5 gives $\int_0^\infty L(x)\mu(x)^j e^{-\alpha\mu(x)}dx = o(\alpha^{\sigma-j})$ hence, by Markov inequality

$$R_{\alpha j} = o(\alpha^{1+\sigma}\ell(\alpha))$$

and $C_{\alpha j}^{(\ell)} \to 0$ in probability as α tends to infinity.

B. Technical Lemma

The proof of the following lemma follows similarly to the proof of Proposition 2 in (Gnedin et al., 2007), and is omitted here.

Lemma B.1 *Let $(X_t)_{t\geq 0}$ be some positive monotone increasing stochastic process with finite first moment $(E(X_t))_{t\geq 0} \in RV_\gamma$ where $\gamma \geq 0$ (see Definition C.1). Assume*

$$\mathrm{var}(X_t) = O\{t^{-a}E(X_t)^2\}$$

for some $a > 0$. Then

$$\frac{X_t}{E(X_t)} \to 1 \text{ almost surely as } t \to \infty.$$

The following lemma is a compilation of results from Propositions 17, 18 and 19 in Gnedin et al. (2007).

Lemma B.2 *Let $\mu : \mathbb{R}_+ \to \mathbb{R}_+$ be a positive, right-continuous and monotone decreasing function with $\int_0^\infty \mu(x)dx < \infty$ and generalised inverse $\mu^{-1}(x) = \inf\{y > 0 \mid f(y) \leq x\}$ satisfying*

$$\mu^{-1}(x) = x^{-\sigma}\ell(1/x) \tag{60}$$

where $\sigma \in [0, 1]$ and ℓ is a slowly varying function. Consider

$$g_0(t) = \int_0^\infty (1 - e^{-t\mu(x)})dx, \quad g_r(t) = \int_0^\infty e^{-t\mu(x)}\mu(x)^r dx. \quad r \geq 1.$$

Then, for any $\sigma \in [0, 1)$

$$g_0(t) \sim \Gamma(1-\sigma)t^\sigma \ell(t) \text{ as } t \to \infty$$

and, for $r \geq 1$,

$$\begin{cases} g_r(t) \sim t^{\sigma-r}\ell(t)\sigma\Gamma(r-\sigma) & \text{if } \sigma \in (0,1) \\ g_r(t) = o\{t^{\sigma-r}\ell(t)\} & \text{if } \sigma = 0 \end{cases}$$

as $t \to \infty$. For $\sigma = 1$, as $t \to \infty$,

$$g_0(t) \sim t\ell_1(t), \quad g_1(t) \sim \ell_1(t), \quad g_r(t) \sim t^{1-r}\ell(t)\Gamma(r-1)$$

where $\ell_1(t) = \int_t^\infty x^{-1}\ell(x)dx$. Note that $\ell(t) = o(\ell_1(t))$ hence $g_r(t) = o\{t^{1-r}\ell_1(t)\}$.

Lemma B.3 *Let $\mu : \mathbb{R}_+ \to \mathbb{R}_+$ be a positive, monotone decreasing function, and $u : \mathbb{R}_+ \to [0, 1]$ a positive and integrable function with $\int_0^\infty u(x)dx < \infty$. Consider $h_0(t) = \int_0^\infty u(x)(1 - e^{-t\mu(x)})dx$ and for $r \geq 1$ $h_r(t) = \int_0^\infty u(x)e^{-t\mu(x)}\mu(x)^r dx$.*
Then, as $t \to \infty$.

$$h_0(t) \sim \int_0^\infty u(x)dx, \quad h_r(t) = o(t^{-r}), \quad r \geq 1.$$

Proof. $h_0(t) \to \int_0^\infty u(x)dx$ by dominated convergence. Using Proposition C.5,

$$\frac{th_1(t)}{\int_0^\infty u(x)dx} \to 0$$

Proceed by induction for the final result. ■

Lemma B.4 *Let μ be a non-negative, non-increasing function on $\mathbb{R}_+$, with $\int_0^\infty \mu(x)dx < \infty$ and such that its generalised inverse μ^{-1} verifies $\mu^{-1}(x) \sim x^{-\sigma}\ell(1/x)$ as $x \to 0$ with $\sigma \in [0, 1]$ and ℓ a slowly varying function. Then as $t \to \infty$, for all $r > \sigma$*

$$\int_{\mathbb{R}_+} \mu(x)^r e^{-t\mu(x)}dx = O\{t^{\sigma-r}\ell(t)\}$$

Proof. Let $r > \sigma$. Let $U(y) = \mu^{-1}(1/y)$. U is non-negative, non-decreasing, with $U(y) \sim y^\sigma \ell(y)$ as $y \to \infty$. Consider the change of variable $x = U(y)$, one obtains

$$\int_0^\infty \{\mu(x)\}^r e^{-t\mu(x)} dx = \int_0^\infty y^{-r} e^{-t/y} dU(y)$$

We follow part of the proof in (Bingham et al., 1987, p.37). Note that $y \to y^{-r} \exp(-t/y)$ is monotone increasing on $[0, t/r]$ and monotone decreasing on $[t/r, \infty)$.

$$\begin{aligned}
\int_0^\infty y^{-r} e^{-t/y} dU(y) &= \left\{ \int_0^{t/r} + \sum_{n=1}^\infty \int_{2^{n-1}t/r}^{2^n t/r} \right\} y^{-r} e^{-t/y} dU(y) \\
&\leq t^{-r} e^{-r} r^r U(t/r) + t^{-r} r^r \sum_{n=1}^\infty 2^{-r(n-1)} U\left(2^n t/r\right) \\
&\leq 2t^{\sigma-r} e^{-r} r^r \ell(t/r) + 2t^{-r} r^r \sum_{n=1}^\infty 2^{-r(n-1)} (2^n t/r)^\sigma \ell\left(2^n t/r\right) \\
&\leq 2t^{\sigma-r} e^{-r} r^r \ell(t/r) + 2^{r+1} t^{\sigma-r} r^{r-\sigma} \sum_{n=1}^\infty 2^{-n(r-\sigma)} \ell\left(2^n t/r\right)
\end{aligned}$$

for t large, using the regular variation property of U. Using Potter's bound (Bingham et al., 1987, Theorem 1.5.6), we have, for any $\delta > 0$ and for t large

$$\ell(2^n t/r) \leq 2\ell(t) \max(1, 2^{n\delta}/r^\delta).$$

Hence, for t large,

$$\int_0^\infty y^{-r} e^{-t/y} dU(y) \lesssim t^{\sigma-r} \ell(t) \left(1 + \sum_{n=1}^\infty 2^{-n(r-\sigma)} \max(r^\delta, 2^{n\delta})\right)$$

Taking $0 < \delta < \frac{r-\sigma}{2}$, the series in the right handside converges. ■

The next lemma is a slight variation of Lemma B.2, with the addition of a slowly varying function in the integrals. Note that the case $\sigma = 0$ and ℓ tends to a constant is not covered.

Lemma B.5 *Let $f : \mathbb{R}_+ \to \mathbb{R}_+$ be a positive, right-continuous and monotone decreasing function with $\int_0^\infty f(x)dx < \infty$ and generalised inverse $f^{-1}(x) = \inf\{y > 0 \mid f(y) \leq x\}$ satisfying*

$$f^{-1}(x) = x^{-\sigma} \ell(1/x) \tag{61}$$

where $\sigma \in [0, 1]$ and ℓ is a slowly varying function, with $\lim_{t\to\infty} \ell(t) = \infty$ if $\sigma = 0$. Consider

$$\widetilde{g}_0(t) = \int_0^\infty (1 - e^{-tf(x)}) L(x) dx$$

and for $r \geq 1$

$$\widetilde{g}_r(t) = \int_0^\infty e^{-tf(x)} f(x)^r L(x) dx.$$

where $L : \mathbb{R}_+ \to (0,\infty)$ is a locally integrable function with $\lim_{t\to\infty} L(t) = b \in [0,\infty)$. Then, for any $\sigma \in [0,1)$

$$\begin{cases} \widetilde{g}_0(t) \sim b\Gamma(1-\sigma)t^\sigma \ell(t) & \text{if } b > 0 \\ \widetilde{g}_0(t) = o(t^\sigma \ell(t)) & \text{if } b = 0 \end{cases}$$

and, for $r \geq 1$,

$$\begin{cases} \widetilde{g}_r(t) \sim bt^{\sigma-r}\ell(t)\sigma\Gamma(r-\sigma) & \text{if } \sigma \in (0,1), b > 0 \\ \widetilde{g}_r(t) = o\{t^{\sigma-r}\ell(t)\} & \text{if } \sigma = 0 \text{ or } b = 0 \end{cases}$$

as $t \to \infty$. For $\sigma = 1$, $b > 0$, as $t \to \infty$,

$$\widetilde{g}_0(t) \sim bt\ell_1(t), \quad \widetilde{g}_1(t) \sim b\ell_1(t), \quad \widetilde{g}_r(t) \sim bt^{1-r}\ell(t)\Gamma(r-1)$$

and where $\ell_1(t) = \int_t^\infty x^{-1}\ell(x)dx$. Note that $\ell(t) = o(\ell_1(t))$ hence $\widetilde{g}_r(t) = o\{t^{1-r}\ell_1(t)\}$.

Proof. Let $g_0(t) = \int_0^\infty (1-e^{-tf(x)})dx$. Let $\ell_1(t) = \int_t^\infty x^{-1}\ell(x)dx$ and $\ell_\sigma(t) = \Gamma(1-\sigma)\ell(t)$ if $\sigma \in [0,1)$. Using Lemma B.2, we have $g_0(t) \sim t^\sigma \ell_\sigma(t)$ as $t \to \infty$, and in particular $g_0(t) \to \infty$. By dominated convergence, for any $x_0 > 0$ $\int_0^{x_0}(1-e^{-tf(x)})L(x)dx \to \int_0^{x_0} L(x)dx < \infty$ hence $\widetilde{g}_0(t) \sim \int_{x_0}^\infty (1-e^{-tf(x)})L(x)dx$ as $t \to \infty$.
Let $\epsilon > 0$. There is x_0 such that for all $x \geq x_0$, $|L(x) - b| \leq \epsilon$ and so

$$(b-\epsilon)\int_{x_0}^\infty (1-e^{-tf(x)})dx \leq \int_{x_0}^\infty (1-e^{-tf(x)})L(x)dx \leq (b+\epsilon)\int_{x_0}^\infty (1-e^{-tf(x)})dx.$$

Hence by sandwiching

$$\lim_{t\to\infty} \frac{\widetilde{g}_0(t)}{t^\sigma \ell_\sigma(t)} = \lim_{t\to\infty} \frac{\int_{x_0}^\infty (1-e^{-tf(x)})L(x)dx}{t^\sigma \ell_\sigma(t)} \in (b-\epsilon, b+\epsilon).$$

As this is true for any $\epsilon > 0$, we obtain $\widetilde{g}_0(t) \sim bt^\sigma \ell_\sigma(t)$ as $t \to \infty$ if $b > 0$ and $\widetilde{g}_0(t) = o(t^\sigma \ell_\sigma(t))$ if $b = 0$. The asymptotic results for $\widetilde{g}_r(t)$ then follow from Proposition C.5. ■

The following is a corollary of (Willmot, 1990, Theorem 2.1.).

Corollary B.6 *(Willmot, 1990, Theorem 2.1.). Assume that*

$$f(x) \sim \ell(x)x^\alpha e^{-\beta x}$$

where ℓ is a slowly varying, locally bounded function on $(0,\infty)$, $\beta \geq 0$ and $\alpha \in \mathbb{R}$, or $\alpha < -1$ and $\beta = 0$. Then, as $n \to \infty$

$$\int_0^\infty \frac{(\lambda x)^n e^{-\lambda x}}{n!} f(x)dx \sim \frac{\ell(n)}{(\lambda+\beta)^{\alpha+1}} \left(\frac{\lambda}{\lambda+\beta}\right)^n n^\alpha \tag{62}$$

and

$$\int_0^\infty \frac{(\lambda x)^n e^{-\lambda x}}{n!} u(x)f(x)dx = o\left(\frac{\ell(n)}{(\lambda+\beta)^{\alpha+1}} \left(\frac{\lambda}{\lambda+\beta}\right)^n n^\alpha\right) \tag{63}$$

for any locally bounded function u vanishing at infinity.

Proof. Equation (62) is proved in (Willmot, 1990, Theorem 2.1.). For any $x_0 > 0$, we have $\int_0^\infty \frac{(\lambda x)^n e^{-\lambda x}}{n!} u(x) f(x) dx \sim \int_{x_0}^\infty \frac{(\lambda x)^n e^{-\lambda x}}{n!} u(x) f(x) dx$. For any $\epsilon > 0$, there is x_0 such that $u(x) < \epsilon$ for all $x > x_0$, hence

$$\int_{x_0}^\infty \frac{(\lambda x)^n e^{-\lambda x}}{n!} u(x) f(x) dx \leq \epsilon \int_0^\infty \frac{(\lambda x)^n e^{-\lambda x}}{n!} f(x) dx$$

and (63) follows from (62) by sandwiching. ■

The following lemma is useful to bound the variance and for the proof of the central limit theorem.

Lemma B.7 *Assume the functions μ and ν satisfy Assumptions 1 and 2, for some $\sigma \in [0, 1]$, slowly varying function ℓ and some $a > \min(1/2, \sigma)$ if $\sigma < 1$ and $a = 1$ if $\sigma = 1$. Then*

$$\int_{\mathbb{R}_+^2} \nu(x, y) e^{-\alpha\mu(x) - \alpha\mu(y) + \alpha\nu(x,y)} dx dy = O\left(\alpha^{2\sigma - 2a} \ell_\sigma^2(\alpha)\right)$$

where ℓ_σ is defined in Equation (20). If $a = 1$ and $\sigma = 0$ we have the stronger result

$$\int_{\mathbb{R}_+^2} \nu(x, y) e^{-\alpha\mu(x) - \alpha\mu(y) + \alpha\nu(x,y)} dx dy = o\left(\alpha^{-2} \ell^2(\alpha)\right).$$

Proof. Using $\nu(x, y) \leq \sqrt{\mu(x)\mu(y)} \leq (\mu(x) + \mu(y))/2$ and Assumption 2,

$$\begin{aligned}\int_{\mathbb{R}_+^2} \nu(x, y) e^{-\alpha\mu(x) - \alpha\mu(y) + \alpha\nu(x,y)} dx dy &\leq \int_{\mathbb{R}_+^2} \nu(x, y) e^{-\alpha\mu(x)/2 - \alpha\mu(y)/2} dx dy \\ &\leq C_1 \left(\int_{x_0}^\infty \mu(x)^a e^{-\alpha\mu(x)/2} \right)^2 + 2 \int_0^{x_0} \int_0^\infty \nu(x, y) e^{-\alpha\mu(x)/2 - \alpha\mu(y)/2} dx dy\end{aligned}$$

where $a > \min(1/2, \sigma)$ if $\sigma < 1$ and $a = 1$ if $\sigma = 1$. Using $\int_0^{x_0} \nu(x, y) dx \leq x_0 \mu(y)$, we have if $x_0 > 0$ (otherwise the bound is trivial)

$$\int_0^{x_0} \int_0^\infty \nu(x, y) e^{-\alpha\mu(x)/2 - \alpha\mu(y)/2} dx dy \leq e^{-\alpha\mu(x_0)/2} x_0 \int_0^\infty \mu(y) e^{-\alpha\mu(y)/2} dy.$$

Since $\mu(x_0) > 0$, the RHS is in $o(\alpha^{-p})$ for any $p > 0$. Using Lemma B.4 ($\sigma < 1$) or B.2 ($\sigma = 1$) together with Assumption 1, we therefore obtain

$$\int_{\mathbb{R}_+^2} \nu(x, y) e^{-\alpha\mu(x) - \alpha\mu(y) + \alpha\nu(x,y)} dx dy = O\{\alpha^{2\sigma - 2a} \ell_\sigma^2(\alpha)\}.$$

In the case $\sigma = 0$ and $a = 1$, Lemma B.2 and Assumption 1 give

$$\int_{\mathbb{R}_+^2} \nu(x, y) e^{-\alpha\mu(x) - \alpha\mu(y) + \alpha\nu(x,y)} dx dy = o\left(\alpha^{-2} \ell^2(\alpha)\right).$$

■

C. Background on regular variation and some technical Lemmas about regularly varying functions

Definition C.1 *A measurable function $U : \mathbb{R}_+ \to \mathbb{R}_+$ is regularly varying at ∞ with index $\rho \in \mathbb{R}$ if for $x > 0$, $\lim_{t\to\infty} U(tx)/U(t) = x^\rho$. We note $U \in RV_\rho$. If $\rho = 0$, we call U slowly varying.*

Proposition C.2 *If $U \in RV_\rho$, then there exists a slowly varying function $\ell \in RV_0$ such that*

$$U(x) = x^\rho \ell(x) \tag{64}$$

Definition C.3 *The de Bruijn conjugate $\ell^\#$ of the slowly varying function ℓ, which always exists, is uniquely defined up to asymptotic equivalence (Bingham et al., 1987, Theorem 1.5.13) by*

$$\ell(x)\ell^\#\{x\ell(x)\} \to 1, \quad \ell^\#(x)\ell\{x\ell^\#(x)\} \to 1$$

as $x \to \infty$. Then $(\ell^\#)^\# \sim \ell$. For example, $(\log^a x)^\# \sim \log^{-a} x$ for $a \neq 0$ and $\ell^\#(x) \sim 1/c$ if $\ell(x) \sim c$.

Proposition C.4 *(Resnick, 1987, Proposition 0.8, Chapter 0) If $U \in RV_\rho$, $\rho \in \mathbb{R}$, and the sequences (a_n) and (a'_n) satisfy $0 < a_n \to \infty$, $0 < a'_n \to \infty$ and $a_n \sim ca'_n$ for some $0 < c < \infty$, then*

$$U(a_n) \sim c^\rho U(a'_n) \text{ as } n \to \infty.$$

Proposition C.5 *(Resnick, 1987, Proposition 0.7 p.21) Let $U : \mathbb{R}_+ \to \mathbb{R}_+$ absolutely continuous with density u, so that $U(x) = \int_0^x u(t)dt$. If $U \in RV_\rho$, $\rho \in \mathbb{R}$ and u is monotone, then*

$$\lim_{x\to\infty} \frac{xu(x)}{U(x)} = \rho$$

and if $\rho \neq 0$, then $sign(\rho)u(x) \in RV_{\rho-1}$.

3.1 Supplementary material

On sparsity, power-law and clustering properties of graphex processes: Supplementary Material

François Caron, Francesca Panero and Judith Rousseau

Department of Statistics, University of Oxford

The supplementary material is organised as follows. Section S1 contains proofs of asymptotic bounds on the variances of the number of nodes, number of nodes of a given degree, and number of triangles of nodes with a given degree, as well as the proof of a secondary proposition for the local clustering coefficient. Section S2 contains proofs of secondary propositions for the central limit theorem. For the sake of simplicity, all Sections, Equations, Lemmas, etc., in the Supplementary material here are denoted with a prefix S, to differentiate them from the Sections, Equations, Lemmas, etc., of the main text (Caron et al., 2020).

S1. Proofs of secondary propositions for the variances and clustering coefficients

S1.1. Proof of Proposition 8 on $\mathrm{var}(N_\alpha)$

An application of the Slivnyak-Mecke and Campbell theorems gives

$$\mathrm{var}(N_\alpha) = E(N_\alpha) + 2\alpha^2 \int_0^\infty \mu(x)(1-W(x,x))e^{-\alpha\mu(x)}dx$$
$$+ \alpha^2 \int_{\mathbb{R}_+^2} (e^{\alpha\nu(x,y)} - 1 + W(x,y))(1-W(x,x))(1-W(y,y))(1-W(x,y))e^{-\alpha\mu(x)-\alpha\mu(y)}dxdy.$$

Using the inequality $e^x - 1 \leq xe^x$,

$$\mathrm{var}(N_\alpha) \leq E(N_\alpha) + 2\alpha^2 \int_{\mathbb{R}_+} \mu(x)e^{-\alpha\mu(x)}dx + \alpha^2 \int_{\mathbb{R}_+^2} e^{-\alpha\mu(x)-\alpha\mu(y)} \left\{\alpha\nu(x,y)e^{\alpha\nu(x,y)} + W(x,y)\right\} dxdy$$

Now, using Lemmas B.2 and B.7

$$\int_{\mathbb{R}_+^2} W(x,y)e^{-\alpha\mu(x)-\alpha\mu(y)}dxdy \leq \int_{\mathbb{R}_+} \mu(x)e^{-\alpha\mu(x)}dx = O\left(\alpha^{\sigma-1}\ell_\sigma(\alpha)\right).$$
$$\int_{\mathbb{R}_+^2} \nu(x,y)e^{-\alpha\mu(x)-\alpha\mu(y)+\alpha\nu(x,y)}dxdy = O\left(\alpha^{2\sigma-2a}\ell_\sigma^2(\alpha)\right).$$

It follows that $\mathrm{var}(N_\alpha) = O(\alpha^{3+2\sigma-2a}\ell_\sigma(\alpha)^2)$.

Assume Assumption 1 and 2 are satisfied, with $a = 1$. From the first part of Proposition 8, we have the upper bound $\mathrm{var}(N_\alpha) = O\left(\alpha^{1+2\sigma}\ell_\sigma^2(\alpha)\right)$.
We now derive a lower bound. If $\sigma = 0$, $\mathrm{var}(N_\alpha) \geq E(N_\alpha) \gtrsim \alpha\ell_0(\alpha)$, hence $\mathrm{var}(N_\alpha) \asymp \alpha\ell(\alpha)$. Consider now the case $\sigma > 0$. We have

$$\mathrm{var}(N_\alpha) \geq \alpha^2 \int_{\mathbb{R}_+^2} (e^{\alpha\nu(x,y)} - 1)(1 - W(x,x))(1 - W(y,y))(1 - W(x,y))e^{-\alpha\mu(x)-\alpha\mu(y)}dxdy$$

and using the inequality $e^x - 1 \geq x$ and Assumption 4

$$\begin{aligned}\mathrm{var}(N_\alpha) &\geq \alpha^3 \int_{\mathbb{R}_+^2} \nu(x,y)(1 - W(x,x))(1 - W(y,y))(1 - W(x,y))e^{-\alpha\mu(x)-\alpha\mu(y)}dxdy \\ &\geq C_0\alpha^3 \int_{x_0}^{\infty}\int_{x_0}^{\infty} \mu(x)\mu(y)(1 - W(x,x))(1 - W(y,y))(1 - W(x,y))e^{-\alpha\mu(x)-\alpha\mu(y)}dxdy\end{aligned}$$

Using Lemmas B.2 and B.3, we have

$$\begin{aligned}&\int_{x_0}^{\infty}\int_{x_0}^{\infty} \mu(x)\mu(y)(1 - W(x,x))(1 - W(y,y))(1 - W(x,y))e^{-\alpha\mu(x)-\alpha\mu(y)}dxdy \\ &\sim \int_0^{\infty}\int_0^{\infty} \mu(x)\mu(y)e^{-\alpha\mu(x)-\alpha\mu(y)}dxdy = \left(\int_{\mathbb{R}_+} \mu(x)e^{-\alpha\mu(x)}dx\right)^2 \sim \alpha^{2\sigma-2}\ell_\sigma^2(\alpha).\end{aligned}$$

It follows that, for $\sigma > 0$, $\mathrm{var}(N_\alpha) \gtrsim \alpha^{1+2\sigma}\ell_\sigma^2(\alpha)$. Combining this with the upper bound gives, for all $\sigma \in [0,1]$ $\mathrm{var}(N_\alpha) \asymp \alpha^{1+2\sigma}\ell_\sigma^2(\alpha)$.

S1.2. Proof of proposition 9 on $\mathrm{var}(N_{\alpha,j})$

We have,

$$\begin{aligned}&E(N_{\alpha,j}^2 \mid M) - E(N_{\alpha,j} \mid M) \\ &= \sum_{i_1 \neq i_2} \mathbb{1}_{\theta_{i_1} \leq \alpha}\mathbb{1}_{\theta_{i_2} \leq \alpha}\mathrm{pr}\left\{\sum_k \mathbb{1}_{\theta_k \leq \alpha}Z_{i_1 k} = j \text{ and } \sum_k \mathbb{1}_{\theta_k \leq \alpha}Z_{i_2,k} = j \mid M\right\}. \\ &= \sum_{b \in \{0,1\}^3}\sum_{j_1=0}^{j}\sum_{i_1 \neq i_2} \mathbb{1}_{\theta_{i_1} \leq \alpha}\mathbb{1}_{\theta_{i_2} \leq \alpha} \\ &\quad \times \mathrm{pr}\left\{\sum_k \mathbb{1}_{\theta_k \leq \alpha}Z_{i_1 k} = j \text{ and } \sum_k \mathbb{1}_{\theta_k \leq \alpha}Z_{i_2,k} = j \text{ and } \sum_k \mathbb{1}_{\theta_k \leq \alpha}Z_{i_1 k}Z_{i_2 k} = j - j_1\right. \\ &\qquad \left.\text{and } Z_{i_1 i_1} = b_{11}, Z_{i_1 i_2} = b_{12}, Z_{i_2 i_2} = b_{22} \mid M\right\}\end{aligned}$$

where $b = (b_{11}, b_{12}, b_{22}) \in \{0,1\}^3$. Let A_1, A_2, A_{12} be disjoint subsets of $\mathbb{N}\backslash\{i_1, i_2\}$ such that $|A_{12}| + b_{12} = j - j_1$, $|A_1| + |A_{1,2}| + b_{11} + b_{12} = |A_2| + |A_{1,2}| + b_{22} + b_{12} = j$ respectively

corresponding to the indices of nodes only connected to node i_1, only to node i_2, or to both nodes (i_1, i_2). Let $A = \{i_1, i_2\} \cup A_1 \cup A_2 \cup A_{12}$. We have

$$
\begin{aligned}
&\mathrm{pr}\left\{\sum_k \mathbb{1}_{\theta_k \le \alpha} Z_{i_1 k} = j, \sum_k \mathbb{1}_{\theta_k \le \alpha} Z_{i_2,k} = j, \sum_k \mathbb{1}_{\theta_k \le \alpha} Z_{i_1 k} Z_{i_2 k} = j - j_1, (Z_{i_1 i_1}, Z_{i_1 i_2}, Z_{i_2 i_2}) = b \mid M\right\} \\
&= \sum_{A_1, A_2, A_{12}} \frac{\mathbb{1}_{\theta_{i_1} \le \alpha} \mathbb{1}_{\theta_{i_2} \le \alpha}}{(j - j_1 - b_{12})!(j_1 - b_{11})!(j_1 - b_{22})!} W(\vartheta_{i_1}, \vartheta_{i_1})^{b_{11}} W(\vartheta_{i_2}, \vartheta_{i_2})^{b_{22}} W(\vartheta_{i_1}, \vartheta_{i_2})^{b_{12}} \\
&\quad \times \{1 - W(\vartheta_{i_1}, \vartheta_{i_1})\}^{1 - b_{11}} \{1 - W(\vartheta_{i_2}, \vartheta_{i_2})\}^{1 - b_{22}} \{1 - W(\vartheta_{i_1}, \vartheta_{i_2})\}^{1 - b_{12}} \\
&\quad \times \left[\prod_{k \in A_1} \mathbb{1}_{\theta_k \le \alpha} W(\vartheta_{i_1}, \vartheta_{i_k}) \{1 - W(\vartheta_{i_2}, \vartheta_{i_k})\}\right] \left[\prod_{k \in A_2} \mathbb{1}_{\theta_k \le \alpha} \{1 - W(\vartheta_{i_1}, \vartheta_{i_k})\} W(\vartheta_{i_2}, \vartheta_{i_k})\right] \\
&\quad \left[\prod_{k \in A_{12}} \mathbb{1}_{\theta_k \le \alpha} W(\vartheta_{i_1}, \vartheta_{i_k}) W(\vartheta_{i_2}, \vartheta_{i_k})\right] \exp\left[-\sum_{k \in \mathbb{N} \setminus A} \{g_{\alpha, \vartheta_{i_1}}(\theta_k, \vartheta_k) + g_{\alpha, \vartheta_{i_2}}(\theta_k, \vartheta_k)\}\right]
\end{aligned}
$$

Using the extended Slivnyak-Mecke theorem,

$$
\begin{aligned}
&E(N_{\alpha,j}^2) - E(N_{\alpha,j}) \\
&= \sum_{b \in \{0,1\}^3} \sum_{j_1 = 0}^{j} \frac{\alpha^{2 + j + j_1 - b_{11} - b_{12} - b_{22}}}{(j - j_1 - b_{12})!(j_1 - b_{11})!(j_1 - b_{22})!} \mathbb{1}_{j_1 \ge b_{11}} \mathbb{1}_{j_1 \ge b_{22}} \mathbb{1}_{j_1 \le j - b_{12}} \\
&\quad \times \int_{\mathbb{R}_+^2} \{\mu(x) - \nu(x,y)\}^{j_1 - b_{11}} \{\mu(y) - \nu(x,y)\}^{j_1 - b_{22}} \nu(x,y)^{j - j_1 - b_{12}} e^{-\alpha\mu(x) - \alpha\mu(y) + \alpha\nu(x,y)} \\
&\qquad \times W(x,x)^{b_{11}} W(y,y)^{b_{22}} W(x,y)^{b_{12}} \{1 - W(x,x)\}^{1 - b_{11}} \{1 - W(y,y)\}^{1 - b_{22}} \{1 - W(x,y)\}^{1 - b_{12}} dxdy \\
&\le \sum_{b \in \{0,1\}^3} \sum_{j_1 = 0}^{j} \frac{\alpha^{2 + j + j_1 - b_{11} - b_{12} - b_{22}}}{(j - j_1 - b_{12})!(j_1 - b_{11})!(j_1 - b_{22})!} \mathbb{1}_{j_1 \ge b_{11}} \mathbb{1}_{j_1 \ge b_{22}} \mathbb{1}_{j_1 \le j - b_{12}} \\
&\quad \times \int_{\mathbb{R}_+^2} \mu(x)^{j_1 - b_{11}} \mu(y)^{j_1 - b_{22}} \nu(x,y)^{j - j_1 - b_{12}} e^{-\alpha\mu(x) - \alpha\mu(y) + \alpha\nu(x,y)} \\
&\qquad \times W(x,x)^{b_{11}} W(y,y)^{b_{22}} W(x,y)^{b_{12}} \{1 - W(x,x)\}^{1 - b_{11}} \{1 - W(y,y)\}^{1 - b_{22}} \{1 - W(x,y)\}^{1 - b_{12}} dxdy
\end{aligned}
$$

We will need the following lemma.

Lemma S1 *Let $r \ge 1$, $j_1, j_2 \ge 0$. Define*

$$
I_r := \int_{\mathbb{R}^2} [\alpha\mu(x)]^{j_1} [\alpha\mu(y)]^{j_2} (\alpha\nu(x,y))^r e^{-\alpha\mu(x) - \alpha\mu(y) + \alpha\nu(x,y)} dxdy
$$

for any $p > 0$. Under Assumptions 1 and 2, we have

$$
I_r = O(\alpha^{r - 2ar + 2\sigma} \ell_\sigma^2(\alpha))
$$

for all $r \ge 1$.

Proof. We have, using Assumption 2, that

$$I_r \leq \alpha^r \int_{\mathbb{R}_+^2} [\alpha\mu(x)]^{j_1}[\alpha\mu(y)]^{j_2}\nu(x,y)^r e^{-\alpha\{\mu(x)+\mu(y)\}/2}dxdy$$
$$\leq C_1^r \alpha^{r-2ar}\left(\int_{\mathbb{R}_+}(\alpha\mu(x))^{j_1+ar}e^{-\alpha\mu(x)/2}dx\right)\left(\int_{\mathbb{R}_+}(\alpha\mu(x))^{j_2+ar}e^{-\alpha\mu(x)/2}dx\right)+o(\alpha^{-p}).$$

for any $p>0$. Assumption 1 and Lemmas B.2 ($\sigma=1$) and B.4 ($\sigma\in[0,1)$) imply that

$$I_r = O(\alpha^{r-2ar+2\sigma}\ell_\sigma^2(\alpha))$$

for all $r\geq 1$. ■

It follows

$$E(N_{\alpha,j}^2)-E(N_{\alpha,j}) \lesssim \sum_{b\in\{0,1\}^3}\frac{\alpha^{2+2j-b_{11}-b_{22}-2b_{12}}}{(j-b_{12}-b_{11})!(j-b_{12}-b_{22})!}\mathbb{1}_{j\geq b_{11}+b_{12}}\mathbb{1}_{j\geq b_{22}+b_{12}}$$
$$\times\int_{\mathbb{R}_+^2}\mu(x)^{j-b_{12}-b_{11}}\mu(y)^{j-b_{12}-b_{22}}e^{-\alpha\mu(x)-\alpha\mu(y)+\alpha\nu(x,y)}$$
$$\times W(x,x)^{b_{11}}W(y,y)^{b_{22}}W(x,y)^{b_{12}}$$
$$\times\{1-W(x,x)\}^{1-b_{11}}\{1-W(y,y)\}^{1-b_{22}}\{1-W(x,y)\}^{1-b_{12}}dxdy$$
$$+O\{\alpha^{2+2\sigma+1-2a}\ell_\sigma^2(\alpha)\}.$$

Let V_0 and V_1 respectively denote the sum of terms such that $b_{12}=0$ and $b_{12}=1$ in the above sum. Using the inequality $e^x\leq 1+xe^x$,

$$V_0 = \sum_{b_{11},b_{22}\in\{0,1\}^2}\frac{\alpha^{2+2j-b_{11}-b_{22}}}{(j-b_{11})!(j-b_{22})!}\int_{\mathbb{R}_+^2}\mu(x)^{j-b_{11}}\mu(y)^{j-b_{12}}e^{-\alpha\mu(x)-\alpha\mu(y)+\alpha\nu(x,y)}$$
$$\times W(x,x)^{b_{11}}W(y,y)^{b_{22}}\{1-W(x,x)\}^{1-b_{11}}\{1-W(y,y)\}^{1-b_{22}}\{1-W(x,y)\}dxdy$$
$$=\sum_{b_{11},b_{22}}\frac{\alpha^{2+2j-b_{11}-b_{22}}}{(j-b_{11})!(j-b_{22})!}\int_{\mathbb{R}_+^2}\mu(x)^{j-b_{11}}\mu(y)^{j-b_{12}}e^{-\alpha\mu(x)-\alpha\mu(y)}$$
$$\times W(x,x)^{b_{11}}W(y,y)^{b_{22}}\{1-W(x,x)\}^{1-b_{11}}\{1-W(y,y)\}^{1-b_{22}}dxdy$$
$$+O\left\{\sum_{b_{11},b_{22}}\frac{\alpha^{3+2j-b_{11}-b_{22}}}{(j-b_{11})!(j-b_{22})!}\int_{\mathbb{R}_+^2}\mu(x)^{j-b_{11}}\mu(y)^{j-b_{12}}\nu(x,y)e^{-\alpha\mu(x)-\alpha\mu(y)+\alpha\nu(x,y)}dxdy\right\}$$
$$=\sum_{b_{11},b_{22}}\left\{\frac{\alpha^{1+j-b_{11}}}{(j-b_{11})!}\int_{\mathbb{R}_+}\mu(x)^{j-b_{11}}W(x,x)^{b_{11}}\{1-W(x,x)\}^{1-b_{11}}e^{-\mu(x)}dx\right\}$$
$$\times\left\{\frac{\alpha^{1+j-b_{22}}}{(j-b_{22})!}\int_{\mathbb{R}_+}\mu(y)^{j-b_{22}}W(y,y)^{b_{22}}\{1-W(y,y)\}^{1-b_{22}}e^{-\mu(y)}dy\right\}$$

$$+ O\{\alpha^{2+2\sigma+1-2a}\ell_\sigma^2(\alpha)\} = E(N_{\alpha,j})^2 + O\{\alpha^{2+2\sigma+1-2a}\ell_\sigma^2(\alpha)\}$$

Similarly,

$$\begin{aligned} V_1 &\leq \sum_{b_{11},b_{22}} \frac{\alpha^{2j-b_{11}-b_{22}}\mathbb{1}_{j\geq 1+b_{11}}\mathbb{1}_{j\geq 1+b_{22}}}{(j-1-b_{11})!(j-1-b_{22})!} \int_{\mathbb{R}_+^2} \mu(x)^{j-1-b_{11}}\mu(y)^{j-1-b_{12}} e^{\alpha\nu(x,y)-\alpha\mu(x)-\alpha\mu(y)} W(x,y)dxdy \\ &\leq \sum_{b_{11},b_{22}} \frac{\alpha^{2j-b_{11}-b_{22}}\mathbb{1}_{j\geq 1+b_{11}}\mathbb{1}_{j\geq 1+b_{22}}}{(j-1-b_{11})!(j-1-b_{22})!} \int_{\mathbb{R}_+^2} \mu(x)^{j-1-b_{11}}\mu(y)^{j-1-b_{12}} e^{-\alpha\mu(x)-\alpha\mu(y)} W(x,y)dxdy \\ &\quad + O\{\alpha^{2+2\sigma+1-2a}\ell_\sigma^2(\alpha)\} \end{aligned}$$

For $j_1 \geq 1$ and $j_2 \geq 1$, using Cauchy-Schwarz and Lemma B.4,

$$\begin{aligned} &\int W(x,y)\mu(x)^{j_1}\mu(y)^{j_2} e^{-\alpha\mu(x)-\alpha\mu(y)}dxdy \\ &\leq \int_{\mathbb{R}_+} \mu(x)^{j_1} e^{-\alpha\mu(x)} \left\{\int W(x,y)\mu(y)^{2j_2} e^{-2\alpha\mu(y)}dy\right\}^{1/2} \mu(x)^{1/2}dx \\ &\leq \left\{\int \mu(x)^{j_1+1/2} e^{-\alpha\mu(x)}dx\right\}\left\{\int \mu(y)^{2j_2} e^{-2\alpha\mu(y)}dy\right\}^{1/2} = O\left\{\alpha^{3\sigma/2-j_1-j_2-1/2}\ell_\sigma^{3/2}(\alpha)\right\} \end{aligned}$$

and for $j_1 \geq 0$

$$\begin{aligned} \int_{\mathbb{R}_+^2} \mu(x)^{j_1} e^{-\alpha\mu(x)-\alpha\mu(y)} W(x,y)dxdy &\leq \int_{\mathbb{R}_+^2} \mu(x)^{j_1} e^{-\alpha\mu(x)} W(x,y)dxdy \\ &= \int_{\mathbb{R}_+} \mu(x)^{j_1+1} e^{-\alpha\mu(x)}dx = O\left\{\alpha^{\sigma-j_1-1}\ell_\sigma(\alpha)\right\} \end{aligned}$$

It follows that $V_1 = O\{\alpha^{2+3\sigma/2-1/2}\ell_\sigma^{3/2}(\alpha)\} + O\{\alpha^{1+\sigma}\ell_\sigma^2(\alpha)\} + O\{\alpha^{2+2\sigma+1-2a}\ell_\sigma^2(\alpha)\}$. Combining the upper bounds on V_0 and V_1, we obtain $\mathrm{var}(N_{\alpha,j}) = O(\alpha^{3-2a+2\sigma}\ell_\sigma^2(\alpha))$ and this terminates the proof. In the case $\sigma = 0$ and $a = 1$, one can use Lemma B.2 instead of Lemma B.4 and replace big O by little o in the above bounds, together with the fact that $E(N_{\alpha,j}) = o(\alpha\ell(\alpha))$ and $\ell(t) = O(\ell^2(t))$ if $\sigma = 0$.

S1.3. Proof of Proposition A.1

We first prove the first equality. The proof is similar to that of Proposition 9, given in Section S1.2. For any $j \geq 2$,

$$2R_{\alpha,j} = 2\sum_i T_{\alpha i}\mathbb{1}_{D_{\alpha i}=j}\mathbb{1}_{\theta_i\leq\alpha} = \sum_{i\neq k\neq l} Z_{ik}Z_{il}Z_{kl}\mathbb{1}_{D_{\alpha i}=j}\mathbb{1}_{\theta_i\leq\alpha}.$$

Let $S_{\alpha j} := 4R_{\alpha j}^2$. We have

$$\begin{aligned}
S_{\alpha j} &= \left(\sum_{i \neq k \neq l} Z_{ik} Z_{il} Z_{kl} \mathbb{1}_{D_{\alpha i}=j} \mathbb{1}_{\theta_i \leq \alpha} \right)^2 \\
&= \sum_{i_1 \neq k_1 \neq l_1 \neq i_2 \neq k_2 \neq l_2} Z_{i_1 k_1} Z_{i_1 l_1} Z_{k_1 l_1} Z_{i_2 k_2} Z_{i_2 l_2} Z_{k_2 l_2} \mathbb{1}_{D_{\alpha i_1}=j} \mathbb{1}_{D_{\alpha i_2}=j} \mathbb{1}_{\theta_{i_1} \leq \alpha} \mathbb{1}_{\theta_{i_2} \leq \alpha} \\
&\quad + 2 \sum_{i_1 \neq k_1 \neq l_1 \neq i_2 \neq k_2} Z_{i_1 k_1} Z_{i_1 l_1} Z_{k_1 l_1} Z_{i_2 k_2} Z_{i_2 l_1} Z_{k_2 l_1} \mathbb{1}_{D_{\alpha i_1}=j} \mathbb{1}_{D_{\alpha i_2}=j} \mathbb{1}_{\theta_{i_1} \leq \alpha} \mathbb{1}_{\theta_{i_2} \leq \alpha} \\
&\quad + 2 \sum_{i_1 \neq k_1 \neq l_1 \neq i_2} Z_{i_1 k_1} Z_{i_1 l_1} Z_{k_1 l_1} Z_{i_2 k_1} Z_{i_2 l_1} \mathbb{1}_{D_{\alpha i_1}=j} \mathbb{1}_{D_{\alpha i_2}=j} \mathbb{1}_{\theta_{i_1} \leq \alpha} \mathbb{1}_{\theta_{i_2} \leq \alpha} \\
&\quad + \sum_{i_1 \neq k_1 \neq l_1 \neq k_2 \neq l_2} Z_{i_1 k_1} Z_{i_1 l_1} Z_{k_1 l_1} Z_{i_1 k_2} Z_{i_1 l_2} Z_{k_2 l_2} \mathbb{1}_{D_{\alpha i_1}=j} \mathbb{1}_{\theta_{i_1} \leq \alpha} \\
&\quad + 2 \sum_{i_1 \neq k_1 \neq l_1 \neq k_2} Z_{i_1 k_1} Z_{i_1 l_1} Z_{k_1 l_1} Z_{i_1 k_2} Z_{k_2 l_1} \mathbb{1}_{D_{\alpha i_1}=j} \mathbb{1}_{\theta_{i_1} \leq \alpha} \\
&\quad + 2 \sum_{i_1 \neq k_1 \neq l_1} Z_{i_1 k_1} Z_{i_1 l_1} Z_{k_1 l_1} \mathbb{1}_{D_{\alpha i_1}=j} \mathbb{1}_{\theta_{i_1} \leq \alpha}
\end{aligned} \tag{S1}$$

Note that some of the terms above are equal to 0 if $j \leq 4$. First note that for any $j_0 \leq j$:

$$\sum_{i \neq k_1 \neq \ldots \neq k_{j_0}} \left(\prod_{l=1}^{j_0} Z_{il} \right) \mathbb{1}_{D_{\alpha i}=j} \mathbb{1}_{\theta_i \leq \alpha} \leq \binom{j}{j_0} N_{\alpha j} \tag{S2}$$

Hence the last three terms of the right-handside of (S1) are upper bounded by $C_j N_{\alpha,j}$, for some constant C_j that does not depend on α. Consider now

$$\begin{aligned}
S_{\alpha,j,1} &= \sum_{i_1 \neq k_1 \neq l_1 \neq i_2 \neq k_2 \neq l_2} Z_{i_1 k_1} Z_{i_1 l_1} Z_{k_1 l_1} Z_{i_2 k_2} Z_{i_2 l_2} Z_{k_2 l_2} \mathbb{1}_{D_{\alpha i_1}=j} \mathbb{1}_{D_{\alpha i_2}=j} \mathbb{1}_{\theta_{i_1} \leq \alpha} \mathbb{1}_{\theta_{i_2} \leq \alpha} \\
&= \sum_{j_1=2}^{j} S_{\alpha,j,1,j_1}
\end{aligned}$$

where, for $j_1 = 2, \ldots, j$

$$\begin{aligned}
S_{\alpha,j,1,j_1} &= \sum_{\substack{i_1 \neq k_1 \neq l_1 \\ \neq i_2 \neq k_2 \neq l_2}} Z_{i_1 k_1} Z_{i_1 l_1} Z_{k_1 l_1} Z_{i_2 k_2} Z_{i_2 l_2} Z_{k_2 l_2} \mathbb{1}_{D_{\alpha i_1}=j} \mathbb{1}_{D_{\alpha i_2}=j} \mathbb{1}_{\sum_k Z_{i_1 k} Z_{i_2 k} \mathbb{1}_{\theta_k \leq \alpha} = j - j_1} \mathbb{1}_{\theta_{i_1} \leq \alpha} \mathbb{1}_{\theta_{i_2} \leq \alpha} \\
&= \sum_{b \in \{0,1\}^3} \sum_{\substack{i_1 \neq k_1 \neq l_1 \\ \neq i_2 \neq k_2 \neq l_2}} Z_{i_1 k_1} Z_{i_1 l_1} Z_{k_1 l_1} Z_{i_2 k_2} Z_{i_2 l_2} Z_{k_2 l_2} \\
&\quad \times \mathbb{1}_{D_{\alpha i_1}=j} \mathbb{1}_{D_{\alpha i_2}=j} \mathbb{1}_{\sum_k Z_{i_1 k} Z_{i_2 k} \mathbb{1}_{\theta_k \leq \alpha} = j - j_1} \mathbb{1}_{Z_{i_1 i_1}=b_{11}} \mathbb{1}_{Z_{i_1 i_2}=b_{12}} \mathbb{1}_{Z_{i_2 i_2}=b_{22}} \mathbb{1}_{\theta_{i_1} \leq \alpha} \mathbb{1}_{\theta_{i_2} \leq \alpha}
\end{aligned}$$

where we introduce $b = (b_{11}, b_{12}, b_{22}) \in \{0,1\}^3$ as in Section S1.2. Using the extended Slivnyak-Mecke theorem, for $j_1 = 2, \ldots, j$,

$$
\begin{aligned}
&E(S_{\alpha,j,1,j_1}) \qquad \text{(S3)}\\
&= \sum_{b\in\{0,1\}^3} \frac{\alpha^{2+j+j_1-b_{11}-b_{12}-b_{22}}}{(j-j_1-b_{12})!(j_1-b_{11}-2)!(j_1-b_{22}-2)!} \mathbb{1}_{j_1\le j-b_{12}} \mathbb{1}_{j_1\ge b_{11}} \mathbb{1}_{j_1\ge b_{22}} \\
&\quad \times \int_{\mathbb{R}_+^6} \{\mu(x_1)-\nu(x_1,x_2)\}^{j_1-2-b_{11}} \{\mu(x_2)-\nu(x_1,x_2)\}^{j_1-2-b_{22}} \nu(x_1,x_2)^{j-j_1-b_{12}} e^{-\alpha\mu(x_1)-\alpha\mu(x_2)+\alpha\nu(x_1,x_2)} \\
&\qquad \times W(x_1,y_1)W(x_1,z_1)W(y_1,z_1)W(x_2,y_2)W(x_2,z_2)W(y_2,z_2) \\
&\qquad \times W(x,x)^{b_{11}} W(y,y)^{b_{22}} W(x,y)^{b_{12}} \\
&\qquad \times \{1-W(x,x)\}^{1-b_{11}} \{1-W(y,y)\}^{1-b_{22}} \{1-W(x,y)\}^{1-b_{12}} dx_1 dy_1 dz_1 dx_2 dy_2 dz_2 \\
&\le \sum_{b\in\{0,1\}^3} \frac{\alpha^{2+j+j_1-b_{11}-b_{12}-b_{22}}}{(j-j_1-b_{12})!(j_1-b_{11}-2)!(j_1-b_{22}-2)!} \mathbb{1}_{j_1\le j-b_{12}} \mathbb{1}_{j_1\ge b_{11}} \mathbb{1}_{j_1\ge b_{22}} \\
&\quad \times \int_{\mathbb{R}_+^6} \mu(x_1)^{j_1-2-b_{11}} \mu(x_2)^{j_1-2-b_{22}} \nu(x_1,x_2)^{j-j_1-b_{12}} e^{-\alpha\mu(x)-\alpha\mu(y)+\alpha\nu(x_1,x_2)} \\
&\qquad \times W(x_1,y_1)W(x_1,z_1)W(y_1,z_1)W(x_2,y_2)W(x_2,z_2)W(y_2,z_2) \\
&\qquad \times W(x_1,x_1)^{b_{11}} W(x_2,x_2)^{b_{22}} W(x_1,x_2)^{b_{12}} \\
&\qquad \times \{1-W(x_1,x_1)\}^{1-b_{11}} \{1-W(x_2,x_2)\}^{1-b_{22}} \{1-W(x_1,x_2)\}^{1-b_{12}} dx_1 dy_1 dz_1 dx_2 dy_2 dz_2 \qquad \text{(S4)}
\end{aligned}
$$

For $b_{12} \neq 0$ or $j \neq j_1$, we can bound the terms in the above sum by

$$
\begin{aligned}
&\frac{\alpha^{2+j+j_1-b_{11}-b_{12}-b_{22}}}{(j-j_1-b_{12})!(j_1-b_{11}-2)!(j_1-b_{22}-2)!} \mathbb{1}_{j_1\le j-b_{12}} \mathbb{1}_{j_1\ge b_{11}} \mathbb{1}_{j_1\ge b_{22}} \\
&\quad \times \int_{\mathbb{R}_+^4} \mu(x_1)^{j_1-b_{11}} \mu(x_2)^{j_1-b_{22}} \nu(x_1,x_2)^{j-j_1-b_{12}} e^{-\alpha\mu(x)-\alpha\mu(y)+\alpha\nu(x_1,x_2)} \\
&\qquad \times W(x_1,x_1)^{b_{11}} W(x_2,x_2)^{b_{22}} W(x_1,x_2)^{b_{12}} \\
&\qquad \times \{1-W(x_1,x_1)\}^{1-b_{11}} \{1-W(x_2,x_2)\}^{1-b_{22}} \{1-W(x_1,x_2)\}^{1-b_{12}} dx_1 dx_2 \\
&= O(\alpha^{3+2\sigma-2a} \ell_\sigma(\alpha)^2) \qquad \text{(S5)}
\end{aligned}
$$

using the intermediate results of the proof in Section S1.2.

Consider now the sum of terms such that $b_{12} = 0$ and $j = j_1$ in (S4). Using the inequality $e^x \le 1 + xe^x$, this sum is upper bounded by

$$
\sum_{b_{11},b_{12}} \frac{\alpha^{2+2j-b_{11}-b_{22}}}{(j-b_{11}-2)!(j-b_{22}-2)!} \int_{\mathbb{R}_+^6} \mu(x_1)^{j-2-b_{11}} \mu(x_2)^{j-2-b_{22}} e^{-\alpha\mu(x)-\alpha\mu(y)+\alpha\nu(x_1,x_2)}
$$

$$
\begin{aligned}
&\times W(x_1,y_1)W(x_1,z_1)W(y_1,z_1)W(x_2,y_2)W(x_2,z_2)W(y_2,z_2)\\
&\times W(x_1,x_1)^{b_{11}}W(x_2,x_2)^{b_{22}}\{1-W(x_1,x_1)\}^{1-b_{11}}\{1-W(x_2,x_2)\}^{1-b_{22}}dx_1dy_1dz_1dx_2dy_2dz_2\\
\leq &\sum_{b_{11},b_{12}}\frac{\alpha^{2+2j-b_{11}-b_{22}}}{(j-b_{11}-2)!(j-b_{22}-2)!}\int_{\mathbb{R}_+^6}\mu(x_1)^{j-2-b_{11}}\mu(x_2)^{j-2-b_{22}}e^{-\alpha\mu(x)-\alpha\mu(y)}\\
&\times W(x_1,y_1)W(x_1,z_1)W(y_1,z_1)W(x_2,y_2)W(x_2,z_2)W(y_2,z_2)\\
&\times W(x_1,x_1)^{b_{11}}W(x_2,x_2)^{b_{22}}\{1-W(x_1,x_1)\}^{1-b_{11}}\{1-W(x_2,x_2)\}^{1-b_{22}}dx_1dy_1dz_1dx_2dy_2dz_2\\
&+\sum_{b_{11},b_{12}}\frac{\alpha^{3+2j-b_{11}-b_{22}}}{(j-b_{11}-2)!(j-b_{22}-2)!}\int_{\mathbb{R}_+^6}\mu(x_1)^{j-2-b_{11}}\mu(x_2)^{j-2-b_{22}}\nu(x,y)e^{-\alpha\mu(x)-\alpha\mu(y)+\alpha\nu(x,y)}\\
&\times W(x_1,y_1)W(x_1,z_1)W(y_1,z_1)W(x_2,y_2)W(x_2,z_2)W(y_2,z_2)\\
&\times W(x_1,x_1)^{b_{11}}W(x_2,x_2)^{b_{22}}\times\{1-W(x_1,x_1)\}^{1-b_{11}}\{1-W(x_2,x_2)\}^{1-b_{22}}dx_1dy_1dz_1dx_2dy_2dz_2\\
\leq &\,4E(R_{\alpha,j})^2\\
&+\sum_{b_{11},b_{12}}\frac{\alpha^{3+2j-b_{11}-b_{22}}}{(j-b_{11}-2)!(j-b_{22}-2)!}\int_{\mathbb{R}_+^2}\mu(x_1)^{j-b_{11}}\mu(x_2)^{j-b_{22}}\nu(x,y)e^{-\alpha\mu(x)-\alpha\mu(y)+\alpha\nu(x,y)}\\
&\times W(x_1,x_1)^{b_{11}}W(x_2,x_2)^{b_{22}}\times\{1-W(x_1,x_1)\}^{1-b_{11}}\{1-W(x_2,x_2)\}^{1-b_{22}}dx_1dx_2\\
= &\,4E(R_{\alpha,j})^2+O(\alpha^{3+2\sigma-2a}\ell_\sigma(\alpha)^2)
\end{aligned}
$$

using Lemma S1 in Section S1.2. It follows that

$$E(S_{\alpha,j,1})=E(4R_{\alpha,j})^2+O(\alpha^{3+2\sigma-2a}\ell_\sigma(\alpha)^2) \tag{S6}$$

Consider now

$$S_{\alpha,j,2}=\sum_{i_1\neq k_1\neq l_1\neq i_2\neq k_2}Z_{i_1k_1}Z_{i_1l_1}Z_{k_1l_1}Z_{i_2k_2}Z_{i_2l_1}Z_{k_2l_1}\mathbb{1}_{D_{\alpha i_1}=j}\mathbb{1}_{D_{\alpha i_2}=j}\mathbb{1}_{\theta_{i_1}\leq\alpha}\mathbb{1}_{\theta_{i_2}\leq\alpha}$$

We have similarly

$$E(S_{\alpha,j,2})=O(\alpha^{3+2\sigma-2a}\ell_\sigma(\alpha)^2) \tag{S7}$$

using Lemma S1. Similarly, using Lemma S1, $E(S_{\alpha,j,3})=O(\alpha^{3+2\sigma-2a}\ell_\sigma(\alpha)^2)$.
Combining the above bound with (S6) and (S7), we obtain $\mathrm{var}(R_{\alpha,j})=O(\alpha^{3+2\sigma-2a}\ell_\sigma(\alpha)^2)$.
We now consider the second bound in Proposition A.1. Consider an increasing sequence $\alpha_n\to\infty$ such that $\alpha_{n+1}-\alpha_n=o(\alpha_n)$ as $n\to\infty$. Let $I_{\alpha_n}=\{i,\,\theta_i\leq\alpha_n\}$ and $I_n^c=I_{\alpha_{n+1}}\backslash I_{\alpha_n}$.

For any $j\geq 1$, let

$$\widetilde{R}_{nj}^{(1)}:=\sum_{i\in I_{\alpha_n}}T_{\alpha_{n+1}i}\mathbb{1}_{D_{\alpha_n i}=j}\mathbb{1}_{\sum_{i'\in I_n^c}Z_{ii'}=1}.$$

We have, similarly to Equation (S1)

$$
\begin{aligned}
(\widetilde{R}_{nj}^{(1)})^2=\sum_{i_1,i_2}^{I_{\alpha_n}}\sum_{k_1\neq l_1\neq i_1}^{I_{\alpha_{n+1}}}\sum_{i_2\neq k_2\neq l_2}^{I_{\alpha_{n+1}}}&Z_{i_1k_1}Z_{i_1l_1}Z_{k_1l_1}Z_{i_2k_2}Z_{i_2l_2}Z_{k_2l_2}\\
&\times\mathbb{1}_{D_{\alpha_n i_1}=j}\mathbb{1}_{D_{\alpha_n i_2}=j}\mathbb{1}_{\sum_{i_1'\in I_n^c}Z_{i_1'i_1}=1}\mathbb{1}_{\sum_{i_2'\in I_n^c}Z_{i_2'i_2}=1}.
\end{aligned}
$$

Hence using the same decomposition as (S1) together with the fact that $\mathbb{1}_{\sum_{i'\in I_n^c} Z_{i'i}=1} \leq 1$, we derive the same bounds as (S2), (S5) and (S7) so that

$$E((\widetilde{R}_{nj}^{(1)})^2) \lesssim \alpha_n^{\sigma+1}\ell_\sigma(\alpha_n) + \alpha_n^{3+2\sigma-2a}\ell_\sigma(\alpha_n)^2 + E(\widetilde{S}_{\alpha_n,j,1,j})$$

where, writing $b = (b_{11}, b_{22})$,

$$\widetilde{S}_{\alpha_n,j,1,j} = \sum_{b\in\{0,1\}^2} \sum_{i_1\neq i_2}^{\in I_{\alpha_n}} \sum_{\substack{k_1\neq l_1\\ \neq k_2\neq l_2}}^{\in I_{\alpha_{n+1}}\setminus\{i_1,i_2\}} Z_{i_1k_1}Z_{i_1l_1}Z_{k_1l_1}Z_{i_2k_2}Z_{i_2l_2}Z_{k_2l_2}\mathbb{1}_{D_{\alpha_n i_1}=j}\mathbb{1}_{D_{\alpha_n i_2}=j}$$
$$\times \mathbb{1}_{\sum_{k\in I_{\alpha_{n+1}}} Z_{i_1k}Z_{i_2k}=0}\mathbb{1}_{Z_{i_1i_1}=b_{11}}\mathbb{1}_{Z_{i_2i_2}=b_{22}}\mathbb{1}_{\sum_{i'\in I_n^c} Z_{i'i_1}=1}\mathbb{1}_{\sum_{i'\in I_n^c} Z_{i'i_2}=1},$$

so that $E(\widetilde{S}_{\alpha_n,j,1,j}) = E(\widetilde{R}_{nj}^{(1)})^2$. We thus obtain that $\mathrm{var}(\widetilde{R}_{nj}^{(1)}) = O(\alpha_n^{3+2\sigma-2a}\ell_\sigma(\alpha_n)^2)$.

S1.4. Proof of Lemma A.2

Let $I_n^c = I_{\alpha_{n+1}}\setminus I_{\alpha_n} = \{i \mid \theta_i \in (\alpha_n, \alpha_{n+1}]\}$. First note that $\widetilde{R}_{nj} = \sum_{r=1}^{j} \widetilde{R}_{nr}^{(1)} + \widetilde{R}_{nr}^{(2)} + \widetilde{R}_{nr}^{(3)}$ where

$$\widetilde{R}_{nr}^{(1)} = \sum_{i\in I_{\alpha_n}} T_{\alpha_{n+1}i}\mathbb{1}_{D_{\alpha_n i}=r}\mathbb{1}_{\sum_{i'\in I_n^c} Z_{ii'}=1},$$
$$\widetilde{R}_{nr}^{(2)} = \sum_{i\in I_{\alpha_{n+1}}} T_{\alpha_{n+1}i}\mathbb{1}_{D_{\alpha_n i}=r}\mathbb{1}_{\sum_{i'\in I_n^c} Z_{ii'}\geq 2},$$
$$\widetilde{R}_{nr}^{(3)} = \sum_{i\in I_n^c} T_{\alpha_{n+1}i}\mathbb{1}_{D_{\alpha_n i}=r}\mathbb{1}_{\sum_{i'\in I_n^c} Z_{ii'}=1}.$$

For any $r \leq j$

$$E\left(\widetilde{R}_{nr}^{(2)}\middle| M\right) \leq \sum_{i\in I_{\alpha_{n+1}}} \sum_{l\neq k}^{\in I_{\alpha_n},\neq i} W(\vartheta_i,\vartheta_l)W(\vartheta_i,\vartheta_k)W(\vartheta_l,\vartheta_k)J_n(i,r-2)\mathrm{pr}\left(\sum_{i'\in I_n^c} Z_{ii'}\geq 2|M\right)$$
$$+ 2\sum_{i\in I_{\alpha_{n+1}}} \sum_{l\neq i}^{\in I_{\alpha_n}} \sum_{k\neq i}^{\in I_n^c} W(\vartheta_i,\vartheta_l)W(\vartheta_i,\vartheta_k)W(\vartheta_l,\vartheta_k)J_n(i,r-1)\mathrm{pr}\left(\sum_{i'\in I_n^c} Z_{ii'}\geq 1|M\right)$$
$$+ \sum_{i\in I_{\alpha_{n+1}}} \sum_{l\neq k}^{\in I_n^c} W(\vartheta_i,\vartheta_l)W(\vartheta_i,\vartheta_k)W(\vartheta_l,\vartheta_k)J_n(i,r)$$

where, recalling the definition of $g_{\alpha,x}$ in Equation (50),

$$J_n(i,r) = \sum_{i_1\neq i_2\ldots\neq i_r\neq l\neq k}^{\in I_{\alpha_n}} \left[\prod_{s=1}^{r} W(\vartheta_i,\vartheta_{i_s})\right] e^{-\sum_{s\neq l,k,i_1,\ldots,i_r}^{I_{\alpha_n}} g_{\alpha,\vartheta_i}(\theta_s,\vartheta_s)}.$$

Note that

$$\mathrm{pr}\left(\sum_{i'\in I_n^c} Z_{ii'} \geq 2|M\right) = 1 - e^{-\sum_{s\in I_n^c} g_{\alpha_{n+1}-\alpha_n,\vartheta_i}(\theta_s,\vartheta_s)} - \sum_{i'\in I_n^c} W(\vartheta_i,\vartheta_{i'}) e^{-\sum_{s\neq i'}^{I_n^c} g_{\alpha_{n+1}-\alpha_n,\vartheta_i}(\theta_s,\vartheta_s)}.$$

Using the Slivnyak-Mecke theorem, the inequality $1-e^{-y}-ye^{-y} \leq y^2$ for $y \geq 0$, the condition (28) and Lemma B.5, we obtain

$$E\left(\widetilde{R}_{nr}^{(2)}\right) \lesssim \alpha_{n+1}\alpha_n^r(\alpha_{n+1}-\alpha_n)^2 \int L_2(x)\mu(x)^{r+2}e^{-\alpha_n\mu(x)}dx \lesssim \alpha_{n+1}\alpha_n^{\sigma-2}(\alpha_{n+1}-\alpha_n)^2\ell(\alpha_n),$$

where $L_2(x)$ converges to $b \geq 0$ at infinity. Noting that $(\alpha_{n+1}-\alpha_n)/\alpha_n = O(1/n)$, we obtain $E\left(\widetilde{R}_{nr}^{(2)}\right) \lesssim \alpha_n^{\sigma+1}\ell(\alpha_n)/n^2$. This implies that $\sum_n E\left(\widetilde{R}_{nr}^{(2)}\right)/(\alpha_n^{\sigma+1}\ell(\alpha_n)) < +\infty$ so that, by Markov inequality and Borel-Cantelli lemma, $\widetilde{R}_{nr}^{(2)} = o(\alpha_n^{\sigma+1}\ell(\alpha_n))$ almost surely as n tends to infinity.
We now study

$$\widetilde{R}_{nr}^{(3)} := \sum_{i\in I_n^c} T_{\alpha_{n+1}i}\mathbb{1}_{D_{\alpha_n i}=r}\mathbb{1}_{\sum_{i'\in I_n^c} Z_{ii'}=1}.$$

Similarly to before

$$\begin{aligned} E\left(\widetilde{R}_{nr}^{(3)} \mid M\right) \leq & \sum_{i\in I_n^c}\sum_{l\neq k}^{\in I_{\alpha_n}} W(\vartheta_i,\vartheta_l)W(\vartheta_i,\vartheta_k)W(\vartheta_l,\vartheta_k)J_n(i,r-2)\mathrm{pr}\left(\sum_{i'\in I_n^c} Z_{ii'}=1|M\right) \\ & + 2\sum_{l\in I_n^c}\sum_{k\in I_{\alpha_n}}^{l\neq i} W(\vartheta_i,\vartheta_l)W(\vartheta_i,\vartheta_k)W(\vartheta_l,\vartheta_k)J_n(i,r-1) \end{aligned}$$

so that

$$E\left(\widetilde{R}_{nr}^{(3)}\right) \lesssim (\alpha_{n+1}-\alpha_n)^2\alpha_n^r \int L_3(x)\mu(x)^{r+1}e^{-\alpha_n\mu(x)}dx \lesssim \frac{\alpha_n^{\sigma+1}\ell(\alpha_n)}{n^2},$$

where $L_3(x)$ converges to b and $\widetilde{R}_{nr}^{(3)} = o(\alpha_n^{\sigma+1}\ell(\alpha_n))$ almost surely as n tends to infinity. Finally, we have

$$\begin{aligned} E\left(\widetilde{R}_{nr}^{(1)}|M\right) \lesssim & \sum_{i\neq l\neq k}^{\in I_{\alpha_n}} W(\vartheta_i,\vartheta_l)W(\vartheta_i,\vartheta_k)W(\vartheta_l,\vartheta_k)J_n(i,r-2)\sum_{i'\in I_n^c} W(\vartheta_i,\vartheta_{i'}) \\ & + 2\sum_{i\neq l}^{\in I_{\alpha_n}}\sum_{k\in I_n^c} W(\vartheta_i,\vartheta_l)W(\vartheta_i,\vartheta_k)W(\vartheta_l,\vartheta_k)J_n(i,r-1) \end{aligned}$$

which implies that

$$\begin{aligned} E\left(\frac{\widetilde{R}_{nr}^{(1)}}{\alpha_n^{\sigma+1}\ell_\sigma(\alpha_n)}\right) & \lesssim \frac{\alpha_n^{r+1}(\alpha_{n+1}-\alpha_n)}{\alpha_n^{\sigma+1}\ell_\sigma(\alpha_n)}\int_{\mathbb{R}_+^4} L_1(x)\mu(x)^{r+1}e^{-\alpha_n\mu(x)}dx \\ & = O\left(\frac{\alpha_{n+1}-\alpha_n}{\alpha_n}\right) = o(1). \end{aligned}$$

where $L_1(x)$ converges to b. Moreover, from Proposition A.1

$$\text{var}\left(\frac{\widetilde{R}_{nr}^{(1)}}{\alpha_n^{\sigma+1}\ell_\sigma(\alpha_n)}\right) = O\left(\alpha_n^{1-2a}\right)$$

so that, $\widetilde{R}_{nr}^{(1)} = o(\alpha_n^{1+\sigma}\ell(\alpha_n))$ almost surely. It finally follows that, for any $j \geq 1$, $\widetilde{R}_{nj} = o(\alpha_n^{1+\sigma}\ell(\alpha_n))$ almost surely as n tends to infinity.

S2. Proof of secondary propositions for the Central Limit Theorem

S2.1. Proof of Proposition 20

Let

$$Z_\alpha := N_\alpha - E(N_\alpha \mid M) = \sum_i \mathbb{1}_{\theta_i \leq \alpha}(\mathbb{1}_{D_{\alpha,i}\geq 1} - (1 - e^{-M(g_{\alpha,\vartheta_i})}))$$

where we recall that $g_{\alpha,x}(\theta,\vartheta) = -\log(1 - W(x,\vartheta))\mathbb{1}_{\theta\leq\alpha}$ and

$$e^{-M(g_{\alpha,\vartheta_i})} = e^{-\sum_j -\log(1-W(\vartheta_i,\vartheta_j))\mathbb{1}_{\theta_j\leq\alpha}} = \prod_j (1 - W(\vartheta_i,\vartheta_j))^{\mathbb{1}_{\theta_j\leq\alpha}}$$

We have $E(Z_\alpha \mid M) = 0$ hence $\text{var}(Z_\alpha) = E(Z_\alpha^2)$. Note that

$$\begin{aligned}
Z_\alpha^2 &= Z_\alpha + \sum_{i_1\neq i_2} \mathbb{1}_{\theta_{i_1}\leq\alpha}(\mathbb{1}_{D_{\alpha,i_1}\geq 1} - (1 - e^{-M(g_{\alpha,\vartheta_{i_1}})}))\mathbb{1}_{\theta_{i_2}\leq\alpha}(\mathbb{1}_{D_{\alpha,i_2}\geq 1} - (1 - e^{-M(g_{\alpha,\vartheta_{i_2}})})) \\
&= Z_\alpha + \sum_{i_1\neq i_2} \mathbb{1}_{\theta_{i_1}\leq\alpha}\mathbb{1}_{D_{\alpha,i_1}\geq 1}\mathbb{1}_{\theta_{i_2}\leq\alpha}\mathbb{1}_{D_{\alpha,i_2}\geq 1} - \sum_{i_1\neq i_2} \mathbb{1}_{\theta_{i_1}\leq\alpha}\mathbb{1}_{D_{\alpha,i_1}\geq 1}\mathbb{1}_{\theta_{i_2}\leq\alpha}(1 - e^{-M(g_{\alpha,\vartheta_{i_2}})}) \\
&- \sum_{i_1\neq i_2} \mathbb{1}_{\theta_{i_1}\leq\alpha}(1 - e^{-M(g_{\alpha,\vartheta_{i_1}})}))\mathbb{1}_{\theta_{i_2}\leq\alpha}\mathbb{1}_{D_{\alpha,i_2}\geq 1} + \sum_{i_1\neq i_2} \mathbb{1}_{\theta_{i_1}\leq\alpha}(1 - e^{-M(g_{\alpha,\vartheta_{i_1}})})\mathbb{1}_{\theta_{i_2}\leq\alpha}(1 - e^{-M(g_{\alpha,\vartheta_{i_2}})})
\end{aligned}$$

We have

$$E\left(Z_\alpha^2 \mid M\right) = \sum_{i_1\neq i_2} \mathbb{1}_{\theta_{i_1}\leq\alpha}\mathbb{1}_{\theta_{i_2}\leq\alpha} e^{-M(g_{\alpha,\vartheta_{i_1}})-M(g_{\alpha,\vartheta_{i_2}})}(e^{g_{\alpha,\vartheta_{i_1}}(\theta_{i_2},\vartheta_{i_2})} - 1)$$

Applying the extended Slivnyak-Mecke theorem

$$\begin{aligned}
E\left(Z_\alpha^2\right) &= \alpha^2 \int_{\mathbb{R}_+^2} (1 - W(x,x))(1 - W(y,y))(1 - W(x,y))e^{-\alpha\mu(x)-\alpha\mu(y)+\alpha\nu(x,y)}(1 - (1 - W(x,y)))dxdy \\
&\leq \alpha^2 \int_{\mathbb{R}_+^2} W(x,y)e^{-\alpha\mu(x)-\alpha\mu(y)+\alpha\nu(x,y)}dxdy.
\end{aligned}$$

Using Cauchy-Schwarz inequality, $\nu(x,y) \leq \sqrt{\mu(x)\mu(y)} \leq \frac{1}{2}(\mu(x)+\mu(y))$, and Lemma B.2, we obtain

$$E\left(Z_\alpha^2\right) = \alpha^2 \int_0^\infty \mu(x) e^{-\alpha/2\mu(x)} dx \asymp \alpha^{1+\sigma}\ell(\alpha) = \begin{cases} O(\alpha^{1+\sigma}\ell_\sigma(\alpha)) & \sigma \in [0,1) \\ o(\alpha\ell(\alpha)) & \sigma = 0 \end{cases}.$$

It follows from Markov's inequality that, in probability

$$Z_\alpha = \begin{cases} O(\alpha^{1/2+\sigma/2}\ell_\sigma^{1/2}(\alpha)) & \sigma \in [0,1) \\ o(\alpha^{1/2}\ell^{1/2}(\alpha)) & \sigma = 0 \end{cases}.$$

S2.2. Proof of Proposition 21

Define $M(h_\alpha) = \sum_i \widetilde{Z}_i$ where $\widetilde{Z}_i = h_\alpha(\theta_i, \vartheta_i) = \mathbb{1}_{\theta_i \leq \alpha}\left[1 - (1 - W(\vartheta_i,\vartheta_i))e^{-\alpha\mu(\vartheta_i)}\right]$. Using Campbell's formula

$$E\left(\sum_i \widetilde{Z}_i\right) = \alpha \int_0^\infty (1 - (1 - W(x,x))e^{-\alpha\mu(x)})dx = E(N_\alpha)$$
$$\text{var}\left(\sum_i \widetilde{Z}_i\right) = \alpha \int_0^\infty \left[1 - (1 - W(x,x))e^{-\alpha\mu(x)}\right]^2 dx \leq E(N_\alpha)$$

Noting that $E(N_\alpha) \sim \alpha^{1+\sigma}\Gamma(1-\sigma)\ell(\alpha)$, it follows from Chebyshev's inequality that, in probability,

$$\sum_i \widetilde{Z}_i - E(N_\alpha) = O\left(\sqrt{\text{var}\left(\sum_i \widetilde{Z}_i\right)}\right) = O\left(\alpha^{1/2+\sigma/2}\ell_\sigma^{1/2}(\alpha)\right)$$

If μ has an unbounded support then, under Assumption 1, either $\sigma > 0$ or $\sigma = 0$ and $\ell(t) \to \infty$. In both cases, in probability, $\sum_i \widetilde{Z}_i - E(N_\alpha) = o\left(\alpha^{1/2+\sigma}\ell_\sigma(\alpha)\right)$.

S2.3. Proof of Proposition 22

Let

$$f_\alpha(M) = \sum_i \mathbb{1}_{\theta_i \leq \alpha}\left[(1 - W(\vartheta_i,\vartheta_i))e^{-\alpha\mu(\vartheta_i)} - e^{-M(g_{\alpha,\vartheta_i})}\right]$$

The idea is to use Theorem 1.1 from Last et al. (2016). To do so, define

$$F_\alpha = \frac{f_\alpha(M)}{\sqrt{v_\alpha}} \tag{S8}$$

where $v_\alpha = \mathrm{var}(N_\alpha) \asymp \alpha^{1+2\sigma}\ell_\sigma^2(\alpha)$. Note that $E(F_\alpha) = 0$ and $\mathrm{var}(F_\alpha) = 1$. Consider the difference operator $D_z F_\alpha$ defined by

$$D_z F_\alpha = \frac{1}{\sqrt{v_\alpha}}(f_\alpha(M + \delta_z) - f_\alpha(M))$$

Also

$$\begin{aligned} D^2_{z_1,z_2} F_\alpha = D_{z_2}(D_{z_1} F_\alpha) &= D_{z_2}\left(\frac{1}{\sqrt{v_\alpha}}(f_\alpha(M + \delta_{z_1}) - f_\alpha(M))\right) \\ &= \frac{1}{\sqrt{v_\alpha}}\left(f_\alpha(M + \delta_{z_1} + \delta_{z_2}) - f_\alpha(M + \delta_{z_1}) - f_\alpha(M + \delta_{z_2}) + f_\alpha(M)\right). \end{aligned}$$

Define

$$\begin{aligned} \gamma_{\alpha,1} &:= 2\left(\int_{\mathbb{R}^6_+} \sqrt{\mathbb{E}(D_{z_1}F_\alpha)^2(D_{z_2}F_\alpha)^2}\sqrt{\mathbb{E}(D^2_{z_1,z_3}F_\alpha)^2(D^2_{z_2,z_3}F_\alpha)^2}dz_1dz_2dz_3\right)^{1/2} \\ \gamma_{\alpha,2} &:= \left(\int_{\mathbb{R}^6_+} \mathbb{E}\left[(D^2_{z_1,z_3}F_\alpha)^2(D^2_{z_2,z_3}F_\alpha)^2\right] dz_1dz_2dz_3\right)^{1/2} \\ \gamma_{\alpha,3} &:= \int_{\mathbb{R}^2_+} \mathbb{E}|D_z F_\alpha|^3 dz \end{aligned}$$

We state a corollary of Theorem 1.1 from Last et al. (2016).

Corollary S2 *(Last et al., 2016, Theorem 1.1) If* $\gamma_{\alpha,1}, \gamma_{\alpha,2}, \gamma_{\alpha,3} \to 0$, *then*

$$F_\alpha = \frac{f_\alpha(M)}{\sqrt{v_\alpha}} \to \mathcal{N}(0,1).$$

The rest of the proof aims to showing that $\gamma_{\alpha,1}, \gamma_{\alpha,2}, \gamma_{\alpha,3} \to 0$. The proof is rather lengthy and therefore split in different subsections. We first state a few notations and lemmas that will be useful in the following.

S2.3.1. Definitions and lemmas

The following lemma, obtained with Hölder's inequality, will be used multiple times.

Lemma S3 *For any* $d \geq 1$ *and any* $z_1, \ldots, z_d > 0$,

$$E\left(\prod_{k=1}^d e^{-M(g_{\alpha,z_k})}\right) \leq e^{-\frac{\alpha}{d}\sum_{k=1}^d \mu(z_k)}.$$

Proof. Using Hölder's inequality, for any $d \geq 1$

$$E\left(\prod_{k=1}^d e^{-M(g_{\alpha,z_k})}\right) \leq \prod_{k=1}^d E\left(e^{-dM(g_{\alpha,z_k})}\right)^{1/d} = \prod_{k=1}^d e^{-\frac{\alpha}{d}\int_0^\infty (1-[1-W(z_k,y)]^d)dy}$$

$$\leq \prod_{k=1}^{d} e^{-\frac{\alpha}{d}\int_0^\infty (1-[1-W(z_k,y)])dy} = e^{-\frac{\alpha}{d}\sum_{k=1}^d \mu(z_k)}$$

■

For $i, j \geq 0$, let

$$H_{i,j}(x_1,x_2) = \int_{\mathbb{R}_+^2} W(x_1,y)^i W(x_2,y)^j e^{-\frac{\alpha}{4}\mu(y)} dy, \quad H_i(x) = H_{i,0}(x,x). \tag{S9}$$

The following lemma compiles various useful bounds.

Lemma S4 *Assume Assumptions 1 and 5. Then*

- *For all $j \geq 1$ and all $x_1, \ldots, x_{j-1} > 0$, $y_1 > 0$ and $p_1, \ldots, p_j \geq 1$, $\alpha > 0$*

$$\int_0^\infty \left(\int_0^\infty W(y_1,x_j)^{p_j} \left[\prod_{k=1}^{j-1} W(y_1,x_k)^{p_k} \right] e^{-\alpha\mu(y_1)} dy_1 \right)^{1/2} W(y_2,x_j) dx_j$$
$$\leq L(y_2)\mu(y_2) \left(\int L(y_1)^{p_j} \mu(y_1)^{p_j} \left[\prod_{k=1}^{j-1} W(y_1,x_k)^{p_k} \right] e^{-\alpha\mu(y_1)} dy_1 \right)^{1/2}$$

- *For any $x_2 > 0$,*

$$\int H_{1,1}(x_1,x_2)^2 dx_1 \leq \int L(y_1)L(y_2)\mu(y_1)\mu(y_2) W(x_2,y_1)W(x_2,y_2) e^{-\frac{\alpha}{4}(\mu(y_1)+\mu(y_2))} dy_1 dy_2 \tag{S10}$$

- *For any $q = 1, 2, \ldots$*

$$\int_0^\infty H_1(x)^q dx = \int \prod_{i=1}^{q} \int W(x,y_i) e^{-\frac{\alpha}{4}\mu(y_i)} dy_i dx \quad = O(\alpha^{q\sigma-q}\ell_\sigma(\alpha)^q) \tag{S11}$$

- *For any $q \geq 1$, $p \geq 1$,*

$$\int (\alpha H_1(x_1))^{q/2} L(x_1)\mu(x_1)^{p/2} e^{-\frac{\alpha}{4}\mu(x_1)} dx_1 = O(\alpha^{(q+1)\sigma/2-p/2}\ell_\sigma(\alpha)^{(q+1)/2}) \tag{S12}$$

- *If $q \leq 3$,*

$$\int L(x_1)\mu(x_1)^p W(x_1,x_2)^q e^{-\frac{\alpha}{4}\mu(x_1)} dx_1 \leq \mu(x_2)^q L(x_2)^q \left(\int L(x_1)^2 \mu(x_1)^{2p} e^{-\frac{\alpha}{2}\mu(x_1)} dx_1 \right)^2. \tag{S13}$$

Proof. The first inequality comes from Hölder's inequality, together with Assumptions 5 and 1. Also (S10) is a consequence of

$$\int H_{1,1}(x_1,x_2)^2 dx_1 = \int \nu(y_1,y_2) W(x_2,y_1) W(x_2,y_2) e^{-\frac{\alpha}{4}(\mu(y_1)+\mu(y_2))} dy_1 dy_2$$

Under Assumption 5, For any $q = 1, 2, \ldots$, using Assumption 1, 5, and Lemma B.5,

$$\int_0^\infty H_1(x)^q dx = \int \prod_{i=1}^q \int W(x, y_i) e^{-\frac{\alpha}{4}\mu(y_i)} dy_i dx$$
$$\leq \prod_{i=1}^q \int L(y_i)\mu(y_i) e^{-\frac{\alpha}{4}\mu(y_i)} dy_i = O(\alpha^{q\sigma - q}\ell_\sigma(\alpha)^q).$$

Using Hölder's inequality,

$$\int (\alpha H_1(x_1))^{q/2} L(x_1)\mu(x_1)^{p/2} e^{-\frac{\alpha}{4}\mu(x_1)} dx_1 \leq \left(\int (\alpha H_1(x_1))^q dx_1 \int L(x_1)^2 \mu(x_1)^p e^{-\frac{\alpha}{2}\mu(x_1)} dx_1 \right)^{1/2}$$
$$= O(\alpha^{(q+1)\sigma/2 - p/2} \ell_\sigma(\alpha)^{(q+1)/2})$$

which proves (S12). Finally recall that from Lemma B.5 that for any $p \geq 1$,

$$\int L(x_1)\mu(x_1)^p e^{-\frac{\alpha}{4}\mu(x_1)} dx_1 = O(\alpha^{\sigma - p}\ell_\sigma(\alpha)).$$

so that using Hölder and Assumption 5, for any $q \leq 3$

$$\int L(x_1)\mu(x_1)^p W(x_1, x_2)^q e^{-\frac{\alpha}{4}\mu(x_1)} dx_1 \leq \mu(x_2)^q L(x_2)^q \left(\int L(x_1)^2 \mu(x_1)^{2p} e^{-\frac{\alpha}{2}\mu(x_1)} dx_1 \right)^2$$

■

S2.3.2. General bounds

Let $z = (t, x)$. Recall that $g_{\alpha,x}(\theta, \vartheta) = -\log(1 - W(x, \vartheta))\mathbb{1}_{\theta \leq \alpha}$.

$$\sqrt{v_\alpha} \times D_z F_\alpha = \mathbb{1}_{t \leq \alpha}(1 - W(x, x)) \left[e^{-\alpha\mu(x)} - e^{-M(g_{\alpha,x})} \right] + \mathbb{1}_{t \leq \alpha} \sum_i \mathbb{1}_{\theta_i \leq \alpha} W(\vartheta_i, x) e^{-M(g_{\alpha,\vartheta_i})}.$$

We have

$$\sqrt{v_\alpha}\, |D_z F_\alpha| \leq \mathbb{1}_{t \leq \alpha} \left(\left| e^{-\alpha\mu(x)} - e^{-M(g_{\alpha,x})} \right| + \sum_i \mathbb{1}_{\theta_i \leq \alpha} W(\vartheta_i, x) e^{-M(g_{\alpha,\vartheta_i})} \right) \tag{S14}$$

Similarly,

$$\sqrt{v_\alpha} D^2_{z_1,z_2}(F_\alpha) = \mathbb{1}_{t_1,t_2 \leq \alpha} W(x_1, x_2) \left[(1 - W(x_1, x_1)) e^{-M(g_{\alpha,x_1})} + (1 - W(x_2, x_2)) e^{-M(g_{\alpha,x_2})} \right]$$
$$- \mathbb{1}_{t_1,t_2 \leq \alpha} \sum_i \mathbb{1}_{\theta_i \leq \alpha} W(\vartheta_i, x_1) W(\vartheta_i, x_2) e^{-M(g_{\alpha,\vartheta_i})}$$

Note that the above is equal to 0 if $t_1 > \alpha$ or $t_2 > \alpha$. For $t_1, t_2 \leq \alpha$

$$|\sqrt{v_\alpha} \times D^2_{z_1,z_2} F_\alpha|^2 \leq 2W(x_1, x_2)^2 (e^{-M(g_{\alpha,x_1})} + e^{-M(g_{\alpha,x_2})})^2$$

$$
\begin{aligned}
&+ 2\left(\sum_i \mathbb{1}_{\theta_i \le \alpha} W(\vartheta_i, x_1) W(\vartheta_i, x_2) e^{-M(g_{\alpha,\vartheta_i})}\right)^2 \\
&\le 4W(x_1,x_2)^2 (e^{-M(g_{\alpha,x_1})} + e^{-M(g_{\alpha,x_2})}) + 2\sum_i \mathbb{1}_{\theta_i \le \alpha} W(\vartheta_i, x_1)^2 W(\vartheta_i, x_2)^2 e^{-2M(g_{\alpha,\vartheta_i})} \\
&+ 2\sum_{i \ne j} \mathbb{1}_{\theta_i \le \alpha} \mathbb{1}_{\theta_j \le \alpha} W(\vartheta_i, x_1) W(\vartheta_i, x_2) W(\vartheta_j, x_1) W(\vartheta_j, x_2) e^{-M(g_{\alpha,\vartheta_i}) - M(g_{\alpha,\vartheta_j})}
\end{aligned}
$$

$$
\begin{aligned}
&v_\alpha^2 (D^2_{z_1,z_3} F_\alpha)^2 (D^2_{z_2,z_3} F_\alpha)^2 \\
&\le 16 W(x_1,x_3)^2 W(x_2,x_3)^2 (e^{-M(g_{\alpha,x_1})} + e^{-M(g_{\alpha,x_3})})(e^{-M(g_{\alpha,x_2})} + e^{-M(g_{\alpha,x_3})}) \\
&\quad + 8W(x_1,x_3)^2 (e^{-M(g_{\alpha,x_1})} + e^{-M(g_{\alpha,x_3})}) \left(\sum_i \mathbb{1}_{\theta_i \le \alpha} W(\vartheta_i, x_2) W(\vartheta_i, x_3) e^{-M(g_{\alpha,\vartheta_i})}\right)^2 \\
&\quad + 8W(x_2,x_3)^2 (e^{-M(g_{\alpha,x_2})} + e^{-M(g_{\alpha,x_3})}) \left(\sum_i \mathbb{1}_{\theta_i \le \alpha} W(\vartheta_i, x_1) W(\vartheta_i, x_3) e^{-M(g_{\alpha,\vartheta_i})}\right)^2 \\
&\quad + 4 \sum_{i_1,i_2,i_3,i_4} \mathbb{1}_{\theta_{i_1} \le \alpha} \mathbb{1}_{\theta_{i_2} \le \alpha} \mathbb{1}_{\theta_{i_3} \le \alpha} \mathbb{1}_{\theta_{i_4} \le \alpha} W(\vartheta_{i_1}, x_1) W(\vartheta_{i_1}, x_3) W(\vartheta_{i_2}, x_1) W(\vartheta_{i_2}, x_3) \\
&\qquad \times W(\vartheta_{i_3}, x_2) W(\vartheta_{i_3}, x_3) W(\vartheta_{i_4}, x_2) W(\vartheta_{i_4}, x_3) e^{-\sum_{k=1}^4 M(g_{\alpha,\vartheta_{i_k}})}
\end{aligned}
$$

We obtain, using the inequality (S10)

$$
\begin{aligned}
&E\left(v_\alpha^2 (D^2_{z_1,z_3} F)^2 (D^2_{z_2,z_3} F)^2\right) \\
&\le C \times \Big(W(x_1,x_3)^2 W(x_2,x_3)^2 (e^{-\alpha/2\mu(x_1))} + e^{-\alpha/2\mu(x_3)})(e^{-\alpha/2\mu(x_2)} + e^{-\alpha/2\mu(x_3)}) \\
&+ (\alpha^2 H_{1,1}(x_2,x_3)^2 + \alpha H_{2,2}(x_2,x_3)) W(x_1,x_3)^2 (e^{-\alpha/3\mu(x_1)} + e^{-\alpha/3\mu(x_3))}) \\
&+ (\alpha^2 H_{1,1}(x_1,x_3)^2 + \alpha H_{2,2}(x_1,x_3)) W(x_2,x_3)^2 (e^{-\alpha/3\mu(x_2)} + e^{-\alpha/3\mu(x_3))}) \\
&+ A_2(x_1,x_2,x_3)\Big)
\end{aligned}
\tag{S15}
$$

for some constant $C > 0$, where

$$
\begin{aligned}
&A_2(x_1,x_2,x_3) \\
&= E\left(\sum_{i_1,i_2,i_3,i_4} \prod_{\ell=1}^4 W(\vartheta_{i_\ell}, x_3) \mathbb{1}_{\theta_{i_\ell} \le \alpha} W(\vartheta_{i_1}, x_1) W(\vartheta_{i_2}, x_1) W(\vartheta_{i_3}, x_2) W(\vartheta_{i_4}, x_2) e^{-\sum_{k=1}^4 M(g_{\alpha,\vartheta_{i_k}})}\right) \\
&= \alpha^4 \int \prod_{\ell=1}^4 W(y_\ell, x_3) W(y_1, x_1) W(y_2, x_1) W(y_2, x_3) W(y_3, x_2) W(y_4, x_2) e^{-\alpha/4(\sum_{i=1}^4 \mu(y_i))} dy_{1:4} \\
&\quad + \alpha^3 \int (W(y_1,x_1)^2 W(y_1,x_3)^2 W(y_2,x_2) W(y_2,x_3) W(y_3,x_2) W(y_3,x_3) \\
&\qquad + W(y_1,x_2)^2 W(y_1,x_3)^2 W(y_2,x_1) W(y_2,x_3) W(y_3,x_1) W(y_3,x_3)) e^{-\alpha/3(\sum_{i=1}^3 \mu(y_i))} dy_{1:3}
\end{aligned}
$$

$$+\alpha^2\int (W(y_1,x_1)^2W(y_1,x_3)^2W(y_2,x_2)^2W(y_2,x_3)^2$$
$$+W(y_1,x_2)W(y_1,x_1)W(y_1,x_3)^2W(y_2,x_1)W(y_2,x_2)W(y_2,x_3)^2)e^{-\alpha/2(\sum_{i=1}^2\mu(y_i))}dy_{1:2}$$
$$+\alpha\int W(y_1,x_1)^2W(y_1,x_2)^2W(y_1,x_3)^4e^{-\alpha\mu(y_1)}dy_1\Big)$$

S2.3.3. Proof that $\gamma_{\alpha,3}\to 0$

We show here that $\gamma_{\alpha,3}\to 0$, or equivalently $\int E\left|\sqrt{v_\alpha}D_zF\right|^3dz=o(\alpha^{3/2+3\sigma}\ell_\sigma^3(\alpha))$. From Equation (S14) and using the inequality $(a+b)^3\le 4(a^3+b^3)$ for any $a,b\ge 0$, a sufficient condition is

$$\int_0^\infty E\left[\left|e^{-\alpha\mu(x)}-e^{-M(g_{\alpha,x})}\right|^3+\left(\sum_i \mathbb{1}_{\theta_i\le\alpha}W(\vartheta_i,x)e^{-M(g_{\alpha,\vartheta_i})}\right)^3\right]dx=o(\alpha^{1/2+3\sigma}\ell_\sigma^3(\alpha)).$$

We have

$$\int_0^\infty \mathbb{E}\left[\left|e^{-\alpha\mu(x)}-e^{-M(g_{\alpha,x})}\right|^3\right]dx\le 2\int E\left(\left(e^{-\alpha\mu(x)}-e^{-M(g_{\alpha,x})}\right)^2\right)dx$$
$$\le\int(1-e^{-2\alpha\mu(x)})dx=O(\alpha^\sigma\ell_\sigma(\alpha))$$

Also under Assumptions 1 and 5, using Lemma S3

$$\int_0^\infty E\left[\left(\sum_i \mathbb{1}_{\theta_i\le\alpha}W(\vartheta_i,x)e^{-M(g_{\alpha,\vartheta_i})}\right)^3\right]dx\le\int_0^\infty E\left[\sum_{i_1,i_2,i_3}\prod_{\ell=1}^3\mathbb{1}_{\theta_{i_\ell}\le\alpha}W(\vartheta_{i_\ell},x)e^{-M(g_{\alpha,\vartheta_{i_\ell}})}\right]dx$$
$$\le\alpha^3\left(\int L(y)\mu(y)e^{-\alpha\mu(y)/3}dy\right)^3+3\alpha^2\left(\int L(y)^2\mu(y)^2e^{-\alpha\mu(y)/3}dy\right)\left(\int L(y)\mu(y)e^{-\alpha\mu(y)/3}dy\right)$$
$$+\alpha\int L(y)^3\mu(y)^3e^{-\alpha\mu(y)/3}dy=O(\alpha^{3\sigma}\ell_\sigma(\alpha)^3).$$

It follows that $\gamma_{\alpha,3}\to 0$ as $\alpha\to\infty$.

S2.3.4. Proof that $\gamma_{\alpha,2}\to 0$

We now need to show that the integral of the right hand-side of Equation (S15) with respect to x_1,x_2,x_3 is $o(\alpha^{-3}v_\alpha^2)=o(\alpha^{-1+4\sigma}\ell_\sigma^2(\alpha))$. For the first term in the right hand-side of the inequality (S15), we have

$$\int W(x_1,x_3)^2W(x_2,x_3)^2(e^{-\frac{\alpha}{2}\mu(x_1))}+e^{-\frac{\alpha}{2}\mu(x_3)})(e^{-\frac{\alpha}{2}\mu(x_2)}+e^{-\frac{\alpha}{2}\mu(x_3)})dx_1dx_2dx_3$$

$$\leq 3\int W(x_1,x_3)W(x_2,x_3)(e^{-\frac{\alpha}{2}(\mu(x_1)+\mu(x_2))}+e^{-\frac{\alpha}{2}\mu(x_3)})dx_1dx_2dx_3$$
$$\leq 3\int \nu(x_1,x_2)e^{-\frac{\alpha}{2}(\mu(x_1)+\mu(x_2))}dx_1dx_2+3\int \mu(x_3)^2e^{-\frac{\alpha}{2}\mu(x_3)}dx_3$$
$$= O(\alpha^{2\sigma-2}\ell_\sigma^2(\alpha))+O(\alpha^{\sigma-2}\ell_\sigma(\alpha))$$

For the second line (and similarly for the third line) in the RHS of Equation (S15), we have, noting that $H_{2,2}(x_2,x_3)\leq H_{1,1}(x_2,x_3)$

$$\int(\alpha^2H_{1,1}(x_2,x_3)^2+\alpha H_{1,1}(x_2,x_3))W(x_1,x_3)^2(e^{-\alpha/3\mu(x_1)}+e^{-\alpha/3\mu(x_3))})dx_1dx_2dx_3$$
$$\leq \alpha^2\int W(x_1,x_3)^2W(y_1,x_2)W(y_1,x_3)W(y_2,x_2)W(y_2,x_3)$$
$$\times(e^{-\alpha/3(\mu(y_1)+\mu(y_2)+\mu(x_1))}+e^{-\alpha/3(\mu(y_1)+\mu(y_2)+\mu(x_3))})dx_1dx_2dx_3dy_1dy_2$$
$$+\alpha\int W(x_1,x_3)^2W(y_1,x_2)W(y_1,x_3)(e^{-\alpha/3(\mu(y_1)+\mu(x_1))}+e^{-\alpha/3(\mu(y_1)+\mu(x_3))})dx_1dx_2dx_3dy_1$$
$$=\alpha^2\int W(x_1,x_3)^2\nu(y_1,y_2)W(y_1,x_3)W(y_2,x_3)$$
$$\times(e^{-\alpha/3(\mu(y_1)+\mu(y_2)+\mu(x_1))}+e^{-\alpha/3(\mu(y_1)+\mu(y_2)+\mu(x_3))})dx_1dx_3dy_1dy_2$$
$$+\alpha\int W(x_1,x_3)^2\mu(y_1)W(y_1,x_3)(e^{-\alpha/3(\mu(y_1)+\mu(x_1))}+e^{-\alpha/3(\mu(y_1)+\mu(x_3))})dx_1dx_3dy_1$$
$$\leq 2\alpha^2\int L(x_1)^2L(y_1)L(y_2)\mu(x_1)^2\mu(y_1)\mu(y_2)e^{-\alpha/3(\mu(y_1)+\mu(y_2)+\mu(x_1))}dx_1dy_1dy_2$$
$$+\alpha\int\mu(y_1)^2L(y_1)L(x_1)^2\mu(x_1)^2e^{-\alpha/3(\mu(y_1)+\mu(x_1))}dx_1dy_1+\alpha\int\mu(x_3)^2L(x_3)^2\mu(y_1)e^{-\alpha/3(\mu(y_1)+\mu(x_3))}dx_3dy_1$$
$$=O(\alpha^{3\sigma-2}\ell_\sigma(\alpha)^3)$$

using Assumption 5 and Lemma B.5. For the third term in the right-handside of Equation (S15), we obtain

$$\int A_2(x_1,x_2,x_3)dx_1dx_2dx_3$$
$$\leq\alpha^4\left(\int L(y)^2\mu(y)^2e^{-\alpha/4\mu(y)}dy\right)^4+\alpha^3\int L(y_1)^4\mu(y_1)^4e^{-\alpha/3\mu(y_1)}dy_1\left(\int L(y)^2\mu(y)^2e^{-\alpha/3\mu(y)}dy\right)^2$$
$$+\alpha^2\left(\int L(y)^4\mu(y)^4e^{-\alpha/2\mu(y)}dy\right)^2+\alpha\int L(y)^8\mu(y)^8e^{-\alpha\mu(y)}dy=O(\alpha^{4\sigma-4})\ell_\sigma^4(\alpha)$$

It follows that $\gamma_{\alpha,2}\to 0$ as $\alpha\to\infty$.

S2.3.5. Proof that $\gamma_{\alpha,1}\to 0$

For any $x>0$ and any unit-rate Poisson point measure M on $\mathbb{R}_+^2$, denote

$$r_\alpha(x,M)=e^{-\alpha\mu(x)}+e^{-M(g_{\alpha,x})} \tag{S16}$$

For any $z_1 = (t_1, x_1), z_2 = (t_2, x_2)$, if $t_1 > \alpha$ or $t_2 > \alpha$, then $|D_{z_1}(F_\alpha)|^2 |D_{z_2}(F_\alpha)|^2 = 0$. Otherwise, if $t_1, t_2 \leq \alpha$, we have from Equation (S14),

$$
\begin{aligned}
&v_\alpha |D_{z_1}(F_\alpha)|^2 |D_{z_2}(F_\alpha)|^2 \\
&\leq \left(4r_\alpha(x_1, M) + 2 \sum_{i,j} \mathbb{1}_{\theta_i \leq \alpha} \mathbb{1}_{\theta_j \leq \alpha} W(\vartheta_i, x_1) W(\vartheta_j, x_1) e^{-M(g_{\alpha,\vartheta_i}) - M(g_{\alpha,\vartheta_j})} \right) \\
&\quad \times \left(4r_\alpha(x_2, M) + 2 \sum_{i,j} \mathbb{1}_{\theta_i \leq \alpha} \mathbb{1}_{\theta_j \leq \alpha} W(\vartheta_i, x_2) W(\vartheta_j, x_2) e^{-M(g_{\alpha,\vartheta_i}) - M(g_{\alpha,\vartheta_j})} \right) \\
&\leq 16 r_\alpha(x_1, M) r_\alpha(x_2, M) \\
&+ 8 \sum_{i,j} \mathbb{1}_{\theta_i, \theta_j \leq \alpha} \left\{ W(\vartheta_i, x_1) W(\vartheta_j, x_1) r_\alpha(x_2, M) + W(\vartheta_i, x_2) W(\vartheta_j, x_2) r_\alpha(x_1, M) \right) e^{-M(g_{\alpha,\vartheta_i}) - M(g_{\alpha,\vartheta_j})} \\
&+ 4 \sum_{i_1, i_2, i_3, i_4} W(\vartheta_{i_1}, x_1) W(\vartheta_{i_2}, x_1) W(\vartheta_{i_3}, x_2) W(\vartheta_{i_4}, x_2) \prod_{k=1}^{4} \left(\mathbb{1}_{\theta_{i_k} \leq \alpha} e^{-M(g_{\alpha,\vartheta_{i_k}})} \right).
\end{aligned}
$$

Note that, using Campbell theorem, together with Lemma S3

$$
E(r_\alpha(x_1, M) r_\alpha(x_2, M)) \leq e^{-\alpha(\mu(x_1) + \mu(x_2))/2}
$$

It follows that, using the extended Slivnyak-Mecke theorem

$$
\begin{aligned}
&v_\alpha^2 E \left(|D_{z_1}(F_\alpha)|^2 |D_{z_2}(F_\alpha)|^2 \right) \\
&\leq C \Bigg(e^{-\frac{\alpha}{2}(\mu(x_1) + \mu(x_2))} + \alpha \int_0^\infty (W(y, x_1) e^{-\frac{\alpha}{2}\mu(x_2)} + W(y, x_2) e^{-\frac{\alpha}{2}\mu(x_1)}) e^{-\frac{\alpha}{2}\mu(y)} dy \\
&+ \alpha^2 \int_{\mathbb{R}_+^2} \left\{ (W(y_1, x_1) W(y_2, x_1) e^{-\frac{\alpha}{3}\mu(x_2)} + W(y_1, x_2) W(y_2, x_2) e^{-\frac{\alpha}{3}\mu(x_1)} \right\} e^{-\alpha\mu(y_1)/3 - \alpha\mu(y_2)/3} dy_1 dy_2 \\
&+ \alpha^3 \int_{\mathbb{R}_+^3} (W(y_1, x_1)^2 W(y_2, x_2) W(y_3, x_2) + W(y_1, x_2)^2 W(y_2, x_1) W(y_3, x_1)) e^{-\alpha \sum_{k=1}^3 \mu(y_k)/3} dy_1 dy_2 dy_3 \\
&+ \alpha^4 \int_{\mathbb{R}_+^4} W(y_1, x_1) W(y_2, x_1) W(y_3, x_2) W(y_4, x_2) e^{-\alpha \sum_{k=1}^4 \mu(y_k)/4} dy_1 dy_2 dy_3 dy_4 \Bigg) \\
&\leq C \Big(e^{-\frac{\alpha}{3}(\mu(x_1) + \mu(x_2))} + \alpha (H_1(x_1) e^{-\frac{\alpha}{3}\mu(x_2)} + H_1(x_2) e^{-\frac{\alpha}{3}\mu(x_1)}) + \alpha^2 (H_1(x_1)^2 e^{-\frac{\alpha}{3}\mu(x_2)} + H_1(x_2)^2 e^{-\frac{\alpha}{3}\mu(x_1)}) \\
&\quad + \alpha^3 (H_1(x_1)^2 H_{2,0}(x_2) + H_{2,0}(x_1) H_1(x_2)^2) + \alpha^4 H_1(x_1)^2 H_1(x_2)^2 \Big)
\end{aligned}
$$

where $H_{i,j}$ are defined in Equation (S9); therefore, for any $t_1, t_2 \leq \alpha$, using the fact that $H_{2,0} \leq H_1$ and that $\sqrt{\sum_{i=1}^p a_i} \leq \sqrt{p} \sum_{i=1}^p \sqrt{a_i}$

$$
v_\alpha \sqrt{\mathbb{E} |D_{z_1} F_\alpha|^2 |D_{z_2} F_\alpha|^2} \leq \sqrt{6C} \sum_{q_1=0}^{2} \sum_{q_2=0}^{2} (\alpha H_1(x_1))^{q_1/2} (\alpha H_1(x_2))^{q_2/2} e^{-\frac{\alpha}{6}(\mu(x_1)\mathbb{1}_{q_1=0} + \mu(x_2)\mathbb{1}_{q_2=0})}
$$

Additionally, from Equation (S15), we have

$$
\begin{aligned}
&v_\alpha \int \sqrt{E\left((D^2_{z_1,z_3}F_\alpha)^2(D^2_{z_2,z_3}F_\alpha)^2\right)}dx_3 \\
&\leq C \times \Bigg(\underbrace{\int W(x_1,x_3)W(x_2,x_3)(e^{-\frac{\alpha}{4}(\mu(x_1)+\mu(x_2))} + e^{-\alpha\mu(x_3)/4})dx_3}_{B_{\alpha,1}(x_1,x_2)} \\
&\quad + \underbrace{\int (\alpha H_{1,1}(x_2,x_3) + \sqrt{\alpha H_{2,2}(x_2,x_3)})W(x_1,x_3)(e^{-\alpha/6\mu(x_1)} + e^{-\alpha/6\mu(x_3))})dx_3}_{B_{\alpha,2}(x_1,x_2)} \\
&\quad + \int (\alpha H_{1,1}(x_1,x_3) + \sqrt{\alpha H_{2,2}(x_1,x_3)})W(x_2,x_3)(e^{-\alpha/6\mu(x_2)} + e^{-\alpha/6\mu(x_3))})dx_3 \\
&\quad + \underbrace{\int \sqrt{A_2(x_1,x_2,x_3)}dx_3}_{B_{\alpha,3}(x_1,x_2)} \Bigg)
\end{aligned}
$$

for some constant C. To show that $\gamma_{\alpha,1} \to 0$, we aim to show that, for any $q_1, q_2 \in \{0,1,2\}$, and any $k = 1,2,3$

$$
\begin{aligned}
I_{\alpha,k}(q_1,q_2) &:= \int (\alpha H_1(x_1))^{q_1/2}(\alpha H_1(x_2))^{q_2/2} e^{-\frac{\alpha}{6}(\mu(x_1)\mathbb{1}_{q_1=0}+\mu(x_2)\mathbb{1}_{q_2=0})} B_{\alpha,k}(x_1,x_2)dx_1dx_2 \\
&= o(\alpha^{4\sigma-1}\ell_\sigma(\alpha)^4)
\end{aligned}
$$

Consider first

$$
\begin{aligned}
I_{\alpha,1}(q_1,q_2) &= \int (\alpha H_1(x_1))^{q_1/2}(\alpha H_1(x_2))^{q_2/2} e^{-\frac{\alpha}{6}(\mu(x_1)\mathbb{1}_{q_1=0}+\mu(x_2)\mathbb{1}_{q_2=0})} \\
&\qquad \times W(x_1,x_3)W(x_2,x_3)(e^{-\frac{\alpha}{4}(\mu(x_1)+\mu(x_2))} + e^{-\alpha\mu(x_3)/4})dx_1dx_2dx_3 \\
&\leq \int (\alpha H_1(x_1))^{q_1/2} L(x_1)\mu(x_1)e^{-\frac{\alpha}{4}\mu(x_1)}dx_1 \int (\alpha H_1(x_2))^{q_2/2} L(x_2)\mu(x_2)e^{-\frac{\alpha}{4}\mu(x_2)}dx_2 \\
&\quad + \int (\alpha H_1(x_1))^{q_1/2}(\alpha H_1(x_2))^{q_2/2} W(x_1,x_3)W(x_2,x_3)e^{-\alpha\mu(x_3)/4}dx_1dx_2dx_3
\end{aligned}
$$

For $q_1, q_2 \geq 1$, using Hölder's inequality and Assumptions 1 and 5,

$$
\begin{aligned}
&\int (\alpha H_1(x_1))^{q_1/2}(\alpha H_1(x_2))^{q_2/2} W(x_1,x_3)W(x_2,x_3)e^{-\alpha/4\mu(x_3)}dx_1dx_2dx_3 \\
&\leq \Bigg(\int (\alpha H_1(x_1))^{q_1}dx_1 \int L(x_3)^2\mu(x_3)^2 e^{-\alpha\mu(x_3)/4}dx_3 \\
&\qquad \times \int (\alpha H_1(x_2))^{q_2}dx_2 \int L(x_3)^2\mu(x_3)^2 e^{-\alpha\mu(x_3)/4}dx_3 \Bigg)^{1/2}
\end{aligned}
$$

$$= O(\alpha^{(q_1/2+q_2/2+1)\sigma-2}\ell_\sigma(\alpha)^{q_1/2+q_2/2+1})$$

Similarly,

$$\int (\alpha H_1(x_1))^{q_1/2} W(x_1,x_3)W(x_2,x_3)e^{-\alpha\mu(x_3)/4}dx_1dx_2dx_3 = O(\alpha^{(q_1/2+1/2)\sigma-2}\ell_\sigma(\alpha)^{q_1/2+1/2})$$

$$\int W(x_1,x_3)W(x_2,x_3)e^{-\alpha\mu(x_3)/4}dx_1dx_2dx_3 = O(\alpha^{\sigma-2}\ell_\sigma(\alpha))$$

and it follows that for any $q_1, q_2 \in \{0,1,2\}$, $I_{\alpha,1}(q_1,q_2) = O(\alpha^{3\sigma-2}\ell_\sigma(\alpha)^3) = o(\alpha^{-1+4\sigma}\ell_\sigma(\alpha)^2)$ as required. Consider now

$$I_{\alpha,2}(q_1,q_2) = \int (\alpha H_1(x_1))^{q_1/2}(\alpha H_1(x_2))^{q_2/2} e^{-\frac{\alpha}{6}(\mu(x_1)\mathbb{1}_{q_1=0}+\mu(x_2)\mathbb{1}_{q_2=0})}$$
$$\times (\alpha H_{1,1}(x_2,x_3) + \sqrt{\alpha H_{2,2}(x_2,x_3)})W(x_1,x_3)(e^{-\alpha/6\mu(x_1)} + e^{-\alpha/6\mu(x_3))})dx_1dx_2dx_3.$$

We have, using Lemma S4

$$I_{\alpha,2,1}(q_1,q_2) := \int (\alpha H_1(x_1))^{q_1/2}(\alpha H_1(x_2))^{q_2/2} e^{-\frac{\alpha}{6}(\mu(x_1)\mathbb{1}_{q_1=0}+\mu(x_2)\mathbb{1}_{q_2=0})}$$
$$\times \alpha H_{1,1}(x_2,x_3)W(x_1,x_3)e^{-\alpha/6\mu(x_1)}dx_1dx_2dx_3$$
$$\leq \alpha \left(\int (\alpha H_1(x_1))^{q_1/2} e^{-\frac{\alpha}{6}\mu(x_1)\mathbb{1}_{q_1=0}} L(x_1)\mu(x_1)e^{-\alpha/6\mu(x_1)}dx_1\right)$$
$$\times \left(\int (\alpha H_1(x_2))^{q_2/2} e^{-\frac{\alpha}{6}\mu(x_2)\mathbb{1}_{q_2=0}} W(x_2,y)L(y)\mu(y)e^{-\frac{\alpha}{4}\mu(y)}dx_2dy\right)$$

If $q_1 = 0$,

$$\int e^{-\frac{\alpha}{3}\mu(x_1)}L(x_1)\mu(x_1)dx_1 = O(\alpha^{\sigma-1}\ell_\sigma(\alpha))$$

and if $q_1 \geq 1$, using the bound (S12),

$$\int (\alpha H_1(x_1))^{q_1/2} L(x_1)\mu(x_1)e^{-\frac{\alpha}{6}\mu(x_1)}dx_1 = O(\alpha^{(q_1+1)\sigma/2-1}\ell_\alpha(\alpha)^{q_1/2+1/2}) \tag{S17}$$

If $q_2 = 0$,

$$\int e^{-\frac{\alpha}{6}\mu(x_2)}W(x_2,y)L(y)\mu(y)e^{-\frac{\alpha}{4}\mu(y)}dx_2dy \leq \int L(y)\mu(y)^2 e^{-\frac{\alpha}{4}\mu(y)}dy = O(\alpha^{\sigma-2}\ell_\sigma(\alpha))$$

If $q_2 \geq 1$, using Hölder's inequality, the bound (S11) and Assumptions 5 and 1,

$$\int (\alpha H_1(x_2))^{q_2/2} W(x_2,y)L(y)\mu(y)e^{-\frac{\alpha}{4}\mu(y)}dx_2dy$$
$$\leq \left(\int (\alpha H_1(x_2))^{q_2}\right)^{1/2} \int \left(\int W(x_2,y)^2dx_2\right)^{1/2} L(y)\mu(y)e^{-\frac{\alpha}{4}\mu(y)}dy$$

$$\leq \left(\int (\alpha H_1(x_2))^{q_2}\right)^{1/2} \int L(y)^2\mu(y)^2 e^{-\frac{\alpha}{4}\mu(y)}dy = O(\alpha^{(q_2/2+1)\sigma-2}\ell_\sigma(\alpha^{q_2/2+1}))$$

Hence $I_{\alpha,2,1}(q_1,q_2) = o(\alpha^{4\sigma-1}\ell_\sigma^2(\alpha))$. Consider now

$$I_{\alpha,2,2}(q_1,q_2) := \int (\alpha H_1(x_1))^{q_1/2}(\alpha H_1(x_2))^{q_2/2} e^{-\frac{\alpha}{6}(\mu(x_1)\mathbb{1}_{q_1=0}+\mu(x_2)\mathbb{1}_{q_2=0})} \times \alpha H_{1,1}(x_2,x_3)W(x_1,x_3)e^{-\alpha/6\mu(x_3)}dx_1dx_2dx_3$$

If $q_1 = 0$, we obtain the following bound, using the same computations as above,

$$\begin{aligned} I_{\alpha,2,2}(q_1,q_2) &\leq \alpha \int e^{-\frac{\alpha}{6}\mu(x_1)}L(x_1)\mu(x_1)dx_1 \int (\alpha H_1(x_2))^{q_2/2} e^{-\frac{\alpha}{6}\mu(x_2)\mathbb{1}_{q_2=0}} W(x_2,y)L(y)\mu(y)e^{-\frac{\alpha}{4}\mu(y)}dx_2dy \\ &= O(\alpha^{4\sigma-2}\ell_\sigma(\alpha)^{4\sigma}) \end{aligned}$$

If $q_1 > 0$, we have

$$\begin{aligned} I_{\alpha,2,2}(q_1,q_2) &\leq \int \left(\int (\alpha H_1(x_1))^{q_1}dx_1\right)^{1/2} \left(\int W(x_1,x_3)^2dx_1\right)^{1/2} (\alpha H_1(x_2))^{q_2/2} e^{-\frac{\alpha}{6}\mu(x_2)\mathbb{1}_{q_2=0}} \\ &\quad \times \alpha H_{1,1}(x_2,x_3)e^{-\alpha/6\mu(x_3)}dx_2dx_3 \\ &\leq \left(\int (\alpha H_1(x_1))^{q_1}dx_1\right)^{1/2} \int L(x_3)\mu(x_3)(\alpha H_1(x_2))^{q_2/2} e^{-\frac{\alpha}{6}\mu(x_2)\mathbb{1}_{q_2=0}} \\ &\quad \times \alpha H_{1,1}(x_2,x_3)e^{-\alpha/6\mu(x_3)}dx_2dx_3 \end{aligned}$$

If $q_2 = 0$, noting that $H_{1,1}(x_2,x_3) \leq L(x_2)\mu(x_2)L(x_3)\mu(x_3)$, we obtain

$$\begin{aligned} I_{\alpha,2,2}(q_1,q_2) &\leq \alpha \left(\int (\alpha H_1(x_1))^{q_1}dx_1\right)^{1/2} \int L(x_2)\mu(x_2)L(x_3)^2\mu(x_3)^2 e^{-\frac{\alpha}{6}\mu(x_2)}e^{-\alpha/6\mu(x_3)}dx_2dx_3 \\ &= O(\alpha^{(q_1/2+2)\sigma-2}\ell_\sigma(\alpha)^{q_1/2+2}) \end{aligned}$$

If $q_2 > 0$, using Hölder's inequality and the bound (S13),

$$\begin{aligned} I_{\alpha,2,2} \leq \alpha &\left(\int (\alpha H_1(x_1))^{q_1}dx_1\right)^{1/2} \left(\int (\alpha H_1(x_2))^{q_2}dx_2\right)^{1/2} \\ &\times \left(\int L(x_3)^2\mu(x_3)^2 e^{-\alpha/3\mu(x_3)}dx_3\right)^{1/2} \int L(y)^2\mu(y)^2 e^{-\frac{\alpha}{4}\mu(y)}dy = O(\alpha^{7\sigma/2-3}\ell(\alpha)^{7\sigma/2}) \end{aligned}$$

Now, using Hölder and the fact that $H_{2,2} \leq H_{1,1}$, together with (S12),

$$I_{\alpha,2,3}(q_1,q_2) := \int (\alpha H_1(x_1))^{q_1/2}(\alpha H_1(x_2))^{q_2/2} e^{-\frac{\alpha}{6}(\mu(x_1)\mathbb{1}_{q_1=0}+\mu(x_2)\mathbb{1}_{q_2=0})}$$

$$\times (\sqrt{\alpha H_{2,2}(x_2,x_3)})W(x_1,x_3)e^{-\alpha/6\mu(x_1)}dx_1dx_2dx_3$$
$$\leq \sqrt{\alpha}\int (\alpha H_1(x_1))^{q_1/2}L(x_1)\mu(x_1)e^{-\alpha\mu(x_1)/6}dx_1$$
$$\times \int (\alpha H_1(x_2))^{q_2/2}\left(\int \mu(y)W(x_2,y)e^{-\frac{\alpha}{4}\mu(y)}dy\right)^{1/2} e^{-\frac{\alpha}{6}\mu(x_2)\mathbb{1}_{q_2=0}}dx_2$$
$$= O(\alpha^{(3+1/4)\sigma-3/2}\ell_\sigma(\alpha)^{3+1/4})$$

Consider

$$I_{\alpha,2,4}(q_1,q_2) := \int (\alpha H_1(x_1))^{q_1/2}(\alpha H_1(x_2))^{q_2/2}e^{-\frac{\alpha}{6}(\mu(x_1)\mathbb{1}_{q_1=0}+\mu(x_2)\mathbb{1}_{q_2=0})}$$
$$\times \sqrt{\alpha H_{2,2}(x_2,x_3)}W(x_1,x_3)e^{-\alpha/6\mu(x_3)}dx_1dx_2dx_3$$

If $q_1 = 0$, then, using the above computations,

$$I_{\alpha,2,4}(q_1,q_2) \leq \int (\alpha H_1(x_2))^{q_2/2}e^{-\frac{\alpha}{6}(\mu(x_1)+\mu(x_2)\mathbb{1}_{q_2=0})}$$
$$\times \sqrt{\alpha H_{1,1}(x_2,x_3)}W(x_1,x_3)dx_1dx_2dx_3$$
$$= O(\alpha^{(3+1/4)\sigma-3/2}\ell_\sigma(\alpha)^{3+1/4})$$

If $q_1 > 0$ and $q_2 = 0$, noting that $H_{2,2}(x_2,x_3) \leq L(x_2)^2\mu(x_2)^2L(x_3)^2\mu(x_3)^2$ and using Hölder's inequality and Assumptions 5 and 1,

$$I_{\alpha,2,4}(q_1,q_2) \leq \sqrt{\alpha}\left(\int (\alpha H_1(x_1))^{q_1}dx_1\right)^{1/2}\int L(x_3)\mu(x_3)e^{-\frac{\alpha}{6}\mu(x_2)}\sqrt{H_{2,2}(x_2,x_3)}e^{-\alpha/6\mu(x_3)}dx_2dx_3$$
$$\leq \sqrt{\alpha}\left(\int (\alpha H_1(x_1))^{q_1}dx_1\right)^{1/2}\int L(x_3)^2\mu(x_3)^2e^{-\alpha/6\mu(x_3)}dx_3\int e^{-\frac{\alpha}{6}\mu(x_2)}\mu(x_2)L(x_2)dx_2$$
$$= O(\alpha^{3\sigma-3}\ell_\sigma(\alpha)^3)$$

If $q_1, q_2 > 0$, noting that $\int H_{1,1}(x_2,x_3) = \int \mu(y)^2e^{-\frac{\alpha}{4}\mu(y)}dy = O(\alpha^{\sigma-2}\ell_\sigma(\alpha))$,

$$I_{\alpha,2,4}(q_1,q_2)$$
$$\leq \sqrt{\alpha}\left(\int (\alpha H_1(x_1))^{q_1}dx_1\right)^{1/2}\times \int L(x_3)\mu(x_3)(\alpha H_1(x_2))^{q_2/2}\sqrt{H_{1,1}(x_2,x_3)}e^{-\alpha/6\mu(x_3)}dx_2dx_3$$
$$\leq \sqrt{\alpha}\left(\int (\alpha H_1(x_1))^{q_1}dx_1\right)^{1/2}\left(\int (\alpha H_1(x_2))^{q_2}dx_2\right)^{1/2}$$
$$\times \left(\int L(x_3)^2\mu(x_3)^2e^{-\alpha/6\mu(x_3)}dx_3\right)^{1/2}\left(\int H_{1,1}(x_2,x_3)dx_3dx_2\right)^{1/2} = O(\alpha^{3\sigma-3/2}\ell_\sigma(\alpha^{3\sigma}))$$

It follows that

$$I_{\alpha,2}(q_1,q_2) = o(\alpha^{-3}v_\alpha^2). \tag{S18}$$

Finally, consider

$$I_{\alpha,3}(q_1,q_2) = \int (\alpha H_1(x_1))^{q_1/2}(\alpha H_1(x_2))^{q_2/2} e^{-\frac{\alpha}{6}(\mu(x_1)\mathbb{1}_{q_1=0}+\mu(x_2)\mathbb{1}_{q_2=0})}\sqrt{A_2(x_1,x_2,x_3)}dx_1dx_2dx_3$$

where

$$\begin{aligned}
&\sqrt{A_2(x_1,x_2,x_3)} \\
&\leq \alpha^2 H_{1,1}(x_1,x_3)H_{1,1}(x_2,x_3) + \alpha^{3/2}\left(H_{1,1}(x_2,x_3)\sqrt{H_{2,2}(x_1,x_3)} + H_{1,1}(x_1,x_3)\sqrt{H_{2,2}(x_2,x_3)}\right) \\
&+ \alpha\left(\sqrt{H_{2,2}(x_1,x_3)}\sqrt{H_{2,2}(x_2,x_3)}\right. \\
&\qquad \left.+\sqrt{W(y_1,x_2)W(y_1,x_1)W(y_1,x_3)^2W(y_2,x_1)W(y_2,x_2)W(y_2,x_3)^2)e^{-\alpha/2(\sum_{i=1}^2\mu(y_i))}dy_{1:2}}\right) \\
&+ \sqrt{\alpha}\left(\int W(y_1,x_1)^2W(y_1,x_2)^2W(y_1,x_3)^4 e^{-\alpha\mu(y_1)}dy_1\right)^{1/2}
\end{aligned}$$

Using Assumption 5,

$$\int H_{1,1}(x_1,x_3))H_{1,1}(x_2,x_3)dx_3 \quad \leq \int W(x_1,y_1)L(y_1)\mu(y_1)e^{-\frac{\alpha}{4}\mu(y_1)}dy_1 \int W(x_2,y_2)L(y_2)\mu(y_2)e^{-\frac{\alpha}{4}\mu(y_2)}dy_2$$

Therefore, using the asymptotic bounds (S12) and Assumption 1,

$$\begin{aligned}
&I_{\alpha,3,1}(q_1,q_2) \\
&:= \alpha^2\int (\alpha H_1(x_1))^{q_1/2}(\alpha H_1(x_2))^{q_2/2}e^{-\frac{\alpha}{6}(\mu(x_1)\mathbb{1}_{q_1=0}+\mu(x_2)\mathbb{1}_{q_2=0})}H_{1,1}(x_1,x_3)H_{1,1}(x_2,x_3)dx_1dx_2dx_3 \\
&\leq \alpha^2\int (\alpha H_1(x_1))^{q_1/2}(\alpha H_1(x_2))^{q_2/2}H_{1,1}(x_1,x_3)H_{1,1}(x_2,x_3)dx_1dx_2dx_3 \\
&\leq \alpha^2\int (\alpha H_1(x_1))^{q_1/2}L(y_1)\mu(y_1)W(x_1,y_1)e^{-\frac{\alpha}{4}\mu(y_1)}dy_1dx_1 \\
&\quad\times\int (\alpha H_1(x_2))^{q_2/2}L(y_2)\mu(y_2)W(x_2,y_2)e^{-\frac{\alpha}{4}\mu(y_2)}dy_2dx_2
\end{aligned}$$

for any $q_1, q_2 \leq 2$. If $q = 0$,

$$\int (\alpha H_1(x))^{q/2}L(y)\mu(y)W(x,y)e^{-\frac{\alpha}{4}\mu(y)}dydx = \int L(y)\mu(y)^2e^{-\frac{\alpha}{4}\mu(y)}dy = O(\alpha^{\sigma-2}\ell_\sigma(\alpha))$$

If $q > 0$, noting that, using Hölder's inequality and Assumption 3,

$$\int (\alpha H_1(x))^{q/2} W(x,y)dx \leq L(y)\mu(y)\left(\int (\alpha H_1(x))^q dx\right)^{1/2}$$

we have

$$\begin{aligned}\int (\alpha H_1(x))^{q/2} L(y)\mu(y) W(x,y) e^{-\frac{\alpha}{4}\mu(y)} dydx &\leq \left(\int (\alpha H_1(x))^q dx\right)^{1/2} \left(\int L(y)^2\mu(y)^2 e^{-\frac{\alpha}{4}\mu(y)} dy\right)^{1/2} \\ &= O(\alpha^{(q/2+1)\sigma-2}\ell_\sigma(\alpha)^{q/2+1}).\end{aligned}$$

It follows that $I_{\alpha,3,1}(q_1,q_2) = O(\alpha^{4\sigma-2}\ell(\alpha)^{4\sigma})$ for any $q_1, q_2 \in \{0,1,2\}$. Consider now

$$\begin{aligned}I_{\alpha,3,2}(q_1,q_2) :=& \alpha^{3/2}\int (\alpha H_1(x_1))^{q_1/2}(\alpha H_1(x_2))^{q_2/2} e^{-\frac{\alpha}{6}(\mu(x_1)\mathbb{1}_{q_1=0}+\mu(x_2)\mathbb{1}_{q_2=0})} H_{1,1}(x_2,x_3) \\ &\times \left(\int W(y_1,x_1)^2 W(y_1,x_3)^2 e^{-\alpha/3\mu(y_1)} dy_1\right)^{1/2} dx_1dx_2dx_3 \\ \leq& \alpha^{3/2}\int (\alpha H_1(x_1))^{q_1/2}(\alpha H_1(x_2))^{q_2/2} H_{1,1}(x_2,x_3) \\ &\times \left(\int W(y_1,x_1)^2 W(y_1,x_3)^2 e^{-\alpha/3\mu(y_1)} dy_1\right)^{1/2} dx_1dx_2dx_3\end{aligned}$$

Using Lemma S4,

$$\begin{aligned}&\int H_{1,1}(x_2,x_3)\left(\int W(y_1,x_1)^2 W(y_1,x_3)^2 e^{-\alpha/3\mu(y_1)} dy_1\right)^{1/2} dx_3 \\ &\leq \int W(x_2,y_2)L(y_2)\mu(y_2)e^{-\frac{\alpha}{4}\mu(y_2)}dy_2 \times \left(\int L(y_1)^2\mu(y_1)^2 W(y_1,x_1)^2 e^{-\frac{\alpha}{3}\mu(y_1)} dy_1\right)^{1/2}\end{aligned}$$

Therefore

$$\begin{aligned}I_{\alpha,3,2}(q_1,q_2) \leq& \alpha^{3/2}\int (\alpha H_1(x_2))^{q_1/2} W(x_2,y_2)L(y_2)\mu(y_2)e^{-\frac{\alpha}{4}\mu(y_2)}dy_2dx_2 \\ &\times \int (\alpha H_1(x_1))^{q_2/2}\left(\int L(y_1)^2\mu(y_1)^2 W(y_1,x_1)^2 e^{-\frac{\alpha}{3}\mu(y_1)} dy_1\right)^{1/2} dx_1\end{aligned}$$

For $q_1 = 0$,

$$\int L(y_2)\mu(y_2)^2 e^{-\frac{\alpha}{4}\mu(y_2)}dy_2 = O(\alpha^{\sigma-2}\ell_\sigma(\alpha)),$$

while for $q_1 \geq 1$,

$$\int (\alpha H_1(x_2))^{q_1/2} W(x_2,y_2)L(y_2)\mu(y_2)e^{-\frac{\alpha}{4}\mu(y_2)}dy_2dx_2$$

$$\leq \left(\int (\alpha H_1(x_2))^{q_1} dx_2 \times \int W(x_2, y_2)^2 L(y_2)^2 \mu(y_2)^2 e^{-\frac{\alpha}{4}\mu(y_2)} dy_2 dx_2 \right)^{1/2}$$
$$\leq \left(\int (\alpha H_1(x_2))^{q_1} dx_2 \times \int L(y_2)^4 \mu(y_2)^4 e^{-\frac{\alpha}{4}\mu(y_2)} dy_2 dx_2 \right)^{1/2} = O(\alpha^{(q_1/2+1/2)\sigma-2} \ell_\sigma(\alpha)^{q_1+1/2}).$$

Additionally, for $q_2 = 0$, using Hölder's inequality and Assumptions 1 and 5,

$$\int \left(\int L(y_1)^2 \mu(y_1)^2 W(y_1, x_1)^2 e^{-\frac{\alpha}{3}\mu(y_1)} dy_1 \right)^{1/2} dx_1$$
$$\leq \int \left(\int L(y_1)^4 \mu(y_1)^4 e^{-\frac{2\alpha}{3}\mu(y_1)} dy_1 \int W(y_1, x_1)^4 dy_1 \right)^{1/4} dx_1$$
$$\leq \int L(x_1)\mu(x_1) dx_1 \left(\int L(y_1)^4 \mu(y_1)^4 e^{-\frac{2\alpha}{3}\mu(y_1)} dy_1 \right)^{1/4} = O(\alpha^{\sigma/4-1} \ell_\sigma(\alpha)^{1/4})$$

while for $q_2 \geq 1$

$$\int (\alpha H_1(x_1))^{q_2/2} \left(\int L(y_1)^2 \mu(y_1)^2 W(y_1, x_1)^2 e^{-\frac{\alpha}{3}\mu(y_1)} dy_1 \right)^{1/2} dx_1$$
$$\leq \left(\int (\alpha H_1(x_1))^{q_2} dx_1 \int L(y_1)^2 \mu(y_1)^2 W(y_1, x_1)^2 e^{-\frac{\alpha}{3}\mu(y_1)} dy_1 dx_1 \right)^{1/2}$$
$$\leq \left(\int (\alpha H_1(x_1))^{q_2} dx_1 \int L(y_1)^4 \mu(y_1)^4 e^{-\frac{\alpha}{3}\mu(y_1)} dy_1 \right)^{1/2} = O(\alpha^{(q_2+1/2)\sigma-2} \ell_\sigma(\alpha)^{q_2/2+1/2}).$$

Hence, for any $q_1, q_2 \leq 2$, $I_{\alpha,3,2} = O(\alpha^{3\sigma-2}\ell_\sigma(\alpha)^3) = o(\alpha^{-3} v_\alpha^2)$. Consider now

$$I_{\alpha,3,3}(q_1, q_2) := \alpha \int (\alpha H_1(x_1))^{q_1/2} (\alpha H_1(x_2))^{q_2/2} e^{-\frac{\alpha}{6}(\mu(x_1)\mathbb{1}_{q_1=0} + \mu(x_2)\mathbb{1}_{q_2=0})}$$
$$\times \left[\left(\int W(y_1, x_1)^2 W(y_1, x_3)^2 W(y_2, x_2)^2 W(y_2, x_3)^2 e^{-\alpha/2(\sum_{i=1}^2 \mu(y_i))} dy_{1:2} \right)^{1/2} \right.$$
$$\left. + \left(\int W(y_1, x_2) W(y_1, x_1) W(y_1, x_3)^2 W(y_2, x_1) W(y_2, x_2) W(y_2, x_3)^2 e^{-\frac{\alpha}{2(\sum_{i=1}^2 \mu(y_i))}} dy_{1:2} \right)^{1/2} \right]$$
$$dx_1 dx_2 dx_3$$

Using Hölder's inequality,

$$\int \left(\int W(y_1, x_1)^2 W(y_1, x_3)^2 W(y_2, x_2)^2 W(y_2, x_3)^2 e^{-\alpha/2(\sum_{i=1}^2 \mu(y_i))} dy_{1:2} \right)^{1/2} dx_3$$
$$\leq \left(\int W(y_1, x_1)^2 W(y_1, x_3)^2 e^{-\frac{\alpha}{2}\mu(y_1)} dy_1 dx_3 \right)^{1/2} \left(\int W(y_2, x_2)^2 W(y_2, x_3)^2 e^{-\frac{\alpha}{2}\mu(y_2)} dy_2 dx_3 \right)^{1/2}$$

For $q = 0$, using Assumption 5 and 1

$$\int e^{-\frac{\alpha}{6}\mu(x)} \left(\int H_{2,2}(x, x_3) dx_3 \right)^{1/2} dx \leq \int e^{-\frac{\alpha}{6}\mu(x)} \left(\int W(y, x)^2 W(y, x_3) dy dx_3 \right)^{1/2} dx$$
$$= O(\alpha^{\sigma-1} \ell_\sigma(\alpha))$$

Additionally,

$$\left(\int W(y_1, x_1)^2 W(y_1, x_3)^2 e^{-\frac{\alpha}{2}\mu(y_1)} dy_1 dx_3 \right)^{1/2} \left(\int W(y_2, x_2)^2 W(y_2, x_3)^2 e^{-\frac{\alpha}{2}\mu(y_2)} dy_2 dx_3 \right)^{1/2}$$
$$\leq \left(\int L(y_1)^2 \mu(y_1)^2 W(y_1, x_1)^2 e^{-\frac{\alpha}{2}\mu(y_1)} dy_1 \right)^{1/2} \left(\int L(y_2)^2 \mu(y_2)^2 W(y_2, x_2)^2 e^{-\frac{\alpha}{2}\mu(y_2)} dy_2 \right)^{1/2}$$

It follows that, for $q \geq 1$, using Hölder's inequality and Assumptions 5 and 1

$$\int (\alpha H_1(x))^{q/2} \left(\int L(y)^2 \mu(y)^2 W(y, x)^2 e^{-\frac{\alpha}{2}\mu(y)} dy \right)^{1/2} dx$$
$$\leq \left(\int (\alpha H_1(x))^q dx \right)^{1/2} \left(\int L(y)^4 \mu(y)^4 e^{-\frac{\alpha}{2}\mu(y)} dy \right)^{1/2} = O\left(\alpha^{(q/2+1/2)\sigma-2} \ell_\sigma(\alpha)^{(q/2+1/2)} \right)$$

Similarly, for the second term of $I_{\alpha,3,3}(q_1, q_2)$

$$\int \left(\int W(y_1, x_2) W(y_1, x_1) W(y_1, x_3)^2 W(y_2, x_1) W(y_2, x_2) W(y_2, x_3)^2 e^{-\alpha/2(\sum_{i=1}^2 \mu(y_i))} dy_{1:2} \right)^{1/2} dx_3$$
$$\leq \left(\int W(y_1, x_1) W(y_1, x_2) W(y_1, x_3)^2 e^{-\frac{\alpha}{2}\mu(y_1)} dy_1 dx_3 \int W(y_2, x_1) W(y_2, x_2) W(y_2, x_3)^2 e^{-\frac{\alpha}{2}\mu(y_2)} dy_2 dx_3 \right)^{1/2}$$

and, using Hölder's inequality and (S12),

$$\int (\alpha H_1(x_1))^{q_1/2} (\alpha H_1(x_2))^{q_2/2} e^{-\frac{\alpha}{6}(\mu(x_1)\mathbb{1}_{q_1=0} + \mu(x_2)\mathbb{1}_{q_2=0})}$$
$$\left(\int W(y_1, x_1) W(y_1, x_2) W(y_1, x_3)^2 e^{-\frac{\alpha}{2}\mu(y_1)} dy_1 dx_3 \right)^{1/2}$$
$$\times \left(\int W(y_2, x_1) W(y_2, x_2) W(y_2, x_3)^2 e^{-\frac{\alpha}{2}\mu(y_2)} dy_2 dx_3 \right)^{1/2} dx_1 dx_2$$
$$\leq \left(\int (\alpha H_1(x_1))^{q_1} e^{-\frac{\alpha}{6}\mu(x_1)\mathbb{1}_{q_1=0}} W(y_1, x_1) \mu(y_1)^3 L(y_1)^2 e^{-\frac{\alpha}{2}\mu(y_1)} dy_1 dx_1 \right)^{1/2}$$
$$\times \left(\int (\alpha H_1(x_2))^{q_2} e^{-\frac{\alpha}{6}\mu(x_2)\mathbb{1}_{q_2=0}} W(y_2, x_2) \mu(y_2)^3 L(y_2)^2 e^{-\frac{\alpha}{2}\mu(y_2)} dy_2 dx_2 \right)^{1/2}$$

For $q = 0$,

$$\left(\int e^{-\frac{\alpha}{6}\mu(x)} W(y, x) \mu(y)^3 L(y)^2 e^{-\frac{\alpha}{2}\mu(y)} dy dx \right)^{1/2} = O(\alpha^{\sigma/2-2} \ell_\sigma(\alpha)^{1/2})$$

while for $q \geq 1$,

$$\left(\int (\alpha H_1(x))^q W(y,x) \mu(y)^3 L(y)^2 e^{-\frac{\alpha}{2}\mu(y)} dy dx \right)^{1/2} = O(\alpha^{(q/2+1/2)\sigma - 3/2} \ell_\sigma(\alpha)^{q/2+1/2})$$

Combining the above results, we obtain, for any $q_1, q_2 \in \{0,1,2\}$, $I_{\alpha,3,3}(q_1,q_2) = o(\alpha^{-3} v_\alpha^2)$. Consider finally

$$I_{\alpha,3,4}(q_1,q_2) := \int (\alpha H_1(x_1))^{q_1/2} (\alpha H_1(x_2))^{q_2/2} e^{-\frac{\alpha}{6}(\mu(x_1)\mathbb{1}_{q_1=0} + \mu(x_2)\mathbb{1}_{q_2=0})} \times \sqrt{\alpha} \left(\int W(y_1,x_1)^2 W(y_1,x_2)^2 W(y_1,x_3)^4 e^{-\alpha\mu(y_1)} dy_1 \right)^{1/2} dx_1 dx_2 dx_3$$

We have

$$\int \left(\int W(y_1,x_1)^2 W(y_1,x_2)^2 W(y_1,x_3)^4 e^{-\alpha\mu(y_1)} dy_1 \right)^{1/2} dx_3$$
$$\leq \left(\int L(x_3)\mu(x_3) dx_3 \right) \left(\int W(y_1,x_1)^6 e^{-\alpha\mu(y_1)} dy_1 \right)^{1/6} \left(\int W(y_1,x_2)^6 e^{-\alpha\mu(y_1)} dy_1 \right)^{1/6}$$

and, for $q_1 \geq 1$, using Hölder's inequality,

$$\int (\alpha H_1(x_1))^{q_1} \left(\int W(y_1,x_1)^6 e^{-\alpha\mu(y_1)} dy_1 \right)^{1/6} dx_1 = O(\alpha^{(q_1+1/6)\sigma - 1} \ell_\sigma(\alpha)^{q_1+1/6}).$$

For $q_1 = 0$,

$$\int e^{-\frac{\alpha}{6}\mu(x_1)} \left(\int W(y_1,x_1)^6 e^{-\alpha\mu(y_1)} dy_1 \right)^{1/6} dx_1 = O(\alpha^{\sigma-1} \ell_\sigma(\alpha))$$

It follows that, for $q_1, q_2 \leq 2$

$$I_{\alpha,3,4}(q_1,q_2) = O(\alpha^{(4+1/3)\sigma - 3/2} \ell_\sigma(\alpha)^{4+1/3}) = o(\alpha^{-3} v_\alpha^2) \tag{S19}$$
$$I_{\alpha,3}(q_1,q_2) = O(\alpha^{(4+1/3)\sigma - 3/2} \ell_\sigma(\alpha)^{4+1/3}) = o(\alpha^{-3} v_\alpha^2) \tag{S20}$$

and, combining the bounds (S2.3.5), (S18) and (S20), we obtain $\gamma_{\alpha,1} \to 0$.

References

Caron, F., F. Panero, and J. Rousseau (2020). On sparsity, power-law and clustering properties of graphs based of graphex processes. Technical report, University of Oxford.

Last, G., G. Peccati, and M. Schulte (2016). Normal approximation on Poisson spaces: Mehler's formula, second order Poincaré inequalities and stabilization. *Probability theory and related fields 165*(3-4), 667–723.

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

Title of Paper	**On sparsity, power-law and clustering properties of graphex processes**
Publication Status	□Published □ Accepted for Publication ☑□Submitted for Publication □Unpublished and unsubmitted work written in a manuscript style
Publication Details	Joint work with Prof. François Caron (University of Oxford) and Prof. Judith Rousseau (University of Oxford). Submitted to the Advances in Applied Probability journal.

Student Confirmation

Student Name:	Francesca Panero		
Contribution to the Paper	I helped to prove the results on the asymptotic behaviour of the clustering coefficients (section 3.3) and the central limit theorem of subgraph counts (section 4.1). I have helped with the general composition and writing of the paper.		
Signature		Date	19/04/2022

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title:	Professor François Caron		
Supervisor comments			
Signature		Date	19/04/2022

This completed form should be included in the thesis, at the end of the relevant chapter.

Chapter 4

Bayesian nonparametric disclosure risk assessment

Stefano Favaro, Francesca Panero, Tommaso Rigon. "Bayesian nonparametric disclosure risk assessment", Electronic Journal of Statistics, Electron. J. Statist. 15(2), 5626-5651, (2021).

Electronic Journal of Statistics
Vol. 15 (2021) 5626–5651
ISSN: 1935-7524
https://doi.org/10.1214/21-EJS1933

Bayesian nonparametric disclosure risk assessment*

Stefano Favaro†, Francesca Panero and Tommaso Rigon

Department of Economics and Statistics,
University of Torino and Collegio Carlo Alberto,
Corso Unione Sovietica 218/bis, 10134,
Torino, Italy
e-mail: stefano.favaro@unito.it

Department of Statistics,
University of Oxford,
24-29 St Giles', OX1 3LB,
Oxford, United Kingdom
e-mail: francesca.panero@stats.ox.ac.uk

Department of Economics, Management and Statistics,
University of Milano-Bicocca,
Piazza dell'Ateneo Nuovo 1, 20126,
Milano, Italy
e-mail: tommaso.rigon@unimib.it

Abstract: Any decision about the release of microdata for public use is supported by the estimation of measures of disclosure risk, the most popular being the number τ_1 of sample uniques that are also population uniques. In such a context, parametric and nonparametric partition-based models have been shown to have: i) the strength of leading to estimators of τ_1 with desirable features, including ease of implementation, computational efficiency and scalability to massive data; ii) the weakness of producing underestimates of τ_1 in realistic scenarios, with the underestimation getting worse as the tail behaviour of the empirical distribution of microdata gets heavier. To fix this underestimation phenomenon, we propose a Bayesian nonparametric partition-based model that can be tuned to the tail behaviour of the empirical distribution of microdata. Our model relies on the Pitman–Yor process prior, and it leads to a novel estimator of τ_1 with all the desirable features of partition-based estimators and that, in addition, allows to reduce underestimation by tuning a "discount" parameter. We show the effectiveness of our estimator through its application to synthetic data and real data.

MSC2020 subject classifications: Primary 62F15, 62G05.
Keywords and phrases: Bayesian nonparametrics, data confidentiality, Dirichlet process prior, disclosure risk assessment, empirical Bayes, Pitman–Yor process prior.

Received May 2021.

*Stefano Favaro received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement No 817257. Stefano Favaro gratefully acknowledge the financial support from the Italian Ministry of Education, University and Research (MIUR), "Dipartimenti di Eccellenza" grant 2018-2022.
†Also affiliated to IMATI-CNR "Enrico Magenes" (Milan, Italy).

Contents

1. Introduction

Releasing microdata for public use requires a careful assessment of the risk of disclosure (Willenborg and Waal [26]). Consider a microdata sample $(X_1, \ldots, X_n)$ of units (individuals) from a finite population of size $N \geq n$, such that each X_i is a record containing identifying and sensitive information for the i-th unit. Identifying information consists of categorical variables which might match known units of the population. A threat of disclosure results from the possibility that an intruder, who could have personal or public information about the population (e.g. knowing who is included in the sample or using other available datasets), might succeed in identifying an individual through such a match, and hence be able to disclose sensitive information. To quantify disclosure risk, microdata units are partitioned according to a categorical variable that is defined by cross-classifying all identifying variables. That is, units X_i's are partitioned into non-empty cells, with each cell containing individuals with the same combination of values of identifying variables. A risk of disclosure arises from cells in which both sample and population frequencies are small, since the rarer the category the more likely the match is correct. Of special interest are cells with frequency 1 (uniques) since, assuming no errors in matching processes or data sources, for these cells the match is guaranteed to be correct (Bethlehem et al. [2], Skinner et al. [24]). This has motivated inferences on measures of disclosure risk that are functionals of the number of uniques, the most popular being the number τ_1 of sample uniques that are also population uniques. Once an estimate $\hat{\tau}_1$ of τ_1 is obtained, a criterion to understand if the data would incur an

excessive risk in being published is to set a relative risk threshold C and check if the proportion of $\hat{\tau}_1$ with respect to the sample size does not exceed it, i.e. $\hat{\tau}_1/n \leq C$ (Bethlehem et al. [2]). If this is not the case, more care must be used before releasing data, possibly applying other privacy preserving methods.

Over the past three decades, a wide range of parametric and nonparametric approaches, both classical (frequentist) and Bayesian, have been proposed to estimate τ_1. One may identify two main streams in the disclosure risk literature: i) modeling the sole microdata partition by parametric and nonparametric partition-based models (Bethlehem et al. [2], Skinner et al. [24], Fienberg and Makov [11], Samuels [21], Skinner and Elliot [23], Camerlenghi et al. [6]); ii) modeling both the microdata partition and associations among identifying variables by parametric and semiparametric latent class models (Reiter [19], Skinner and Shlomo [25], Manrique-Vallier and Reiter [13, 14], Carota et al. [4, 5]). All these approaches have been applied to synthetic data and real data, showing the effectiveness of τ_1 as a sensible global measure for assessing the risk of disclosure. Partition-based models lead to estimators that are simple, linear in the sampling information, computationally efficient and scalable to massive data sets, though they typically show underestimation when the sampling fraction n/N becomes smaller than a certain threshold (Camerlenghi et al. [6]). Latent class models have typically a better empirical performance than partition-based models, especially for small sampling fractions, though this is achieved at the cost of an increased computational effort for the need of Markov chain Monte Carlo methods for posterior approximation (Reiter [19], Manrique-Vallier and Reiter [13]).

In this paper, we contribute to the partition-based literature from a Bayesian nonparametric perspective. Bayesian nonparametric ideas for estimating τ_1 date back to the seminal work of Samuels [21], where the Dirichlet process (Ferguson [10]) was applied as a prior model for the microdata partition. This approach leads to an estimator of τ_1 which is easy to implement, computationally efficient, and scalable to massive data. Despite these desirable features, empirical analyses in Samuels [21] show that such an approach underestimates τ_1 in many realistic scenarios, the issue being related to the tail behaviour of the empirical distribution of microdata. That is, the heavier the tail the worse the underestimation of τ_1. As heavy-tail scenarios occur when the number of sample uniques is large with respect to the population size, this phenomenon is a critical concern in disclosure risk assessment. A simulation study in Figure 1 shows analogous estimation issues for the most common partition-based estimators of τ_1 in such a heavy-tails setting. Our experiments use synthetic microdata from a power-law distribution of exponent $\sigma > 1$, samples being the 10% of the population of size 10^6, and they are averaged over 1000 iterations. It emerges that the smaller σ, namely the heavier the tail, the worse the underestimation of Bayesian parametric estimators (Bethlehem et al. [2], Skinner et al. [24]), and the worse the overestimation of a nonparametric empirical Bayes estimator (Camerlenghi et al. [6]).

To overcome the underestimation phenomenon of Samuels' approach, we propose a Bayesian nonparametric partition-based model that can be tuned to the

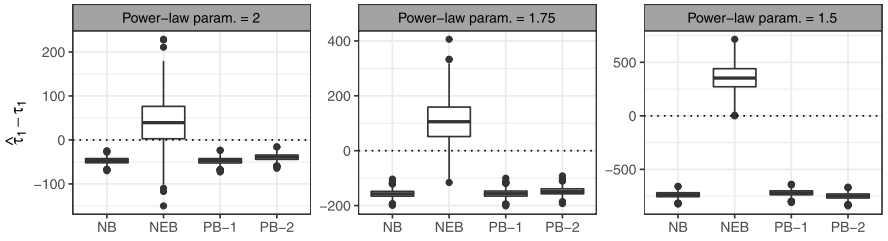


FIG 1. *Empirical performance, with respect to the true* τ_1*, of estimators* $\hat{\tau}_1$*: nonparametric Bayes (*NB*) of Samuels [21], nonparametric empirical Bayes (*NEB*) of Camerlenghi et al. [6], parametric Bayes (*PB*-1) of Bethlehem et al. [2], parametric Bayes (*PB*-2) of Skinner et al. [24].*

tail behaviour of the empirical distribution of microdata. In particular, as a prior model for the microdata partition, we assume the Pitman–Yor process (Perman et al. [15], Pitman [16], Pitman and Yor [18]). The Pitman–Yor process prior generalizes the Dirichlet process prior by means of an additional "discount" parameter that allows to control the tail behaviour of the prior, ranging from geometric tails to heavy power-law tails (Pitman and Yor [18]). Under the Pitman–Yor process prior, we present a simple characterization of the posterior distribution of τ_1, given the observed microdata, and we propose the posterior mean as a Bayesian nonparametric estimator of τ_1. Such an estimator has all the same desirable features as Samuels's estimator and, in addition, it allows to reduce its underestimation of τ_1 by tuning the "discount" parameter with respect to observable microdata. Our approach stands out for being the first partition-based approach to provide a closed-form posterior distribution of τ_1, which makes straightforward to quantify uncertainty of our Bayesian procedure through credible intervals. We investigate the empirical performance of our approach through synthetic data and real data from the 2018 American Community Survey, showing its effectiveness in reducing underestimation phenomenon of Samuels' approach.

The paper is structured as follows. In Section 2 we introduce the Pitman–Yor process prior and its sampling structure, and present our Bayesian nonparametric approach to infer τ_1. Section 3 contains an illustration of the proposed approach through synthetic data and real data. In Section 4 we conclude by discussing our results and directions for future work. Proofs are deferred to the Appendix.

2. Bayesian nonparametric inference for τ_1

We consider a super-population of units belonging to an (ideally) infinite number of distinct symbols $(z_j)_{j\geq 1}$, taking values in a measurable space $\mathbb{Z}$, with unknown proportions $(p_j)_{j\geq 1}$ such that $\sum_{j\geq 1} p_j = 1$. The partition of microdata into non-empty cells, both at the sample and population level, is modeled

as a random partition induced by sampling from the unknown discrete distribution $P = \sum_{j\geq 1} p_j \delta_{z_j}$, where each symbol $z_j \in \mathbb{Z}$ takes the interpretation of a distinct combination of values of identifying variables. That is, a population of $N \geq 1$ of microdata units is assumed to be a random sample $(X_1, \ldots, X_N)$ from P, of which the first $n < N$ elements $(X_1, \ldots, X_n)$ are observable. These samples induce a random partition at population level consisting of K_N cells of frequencies $(N_{1,N}, \ldots, N_{K_N,N})$, and a random partition at the sample level consisting of K_n cells of frequencies $(N_{1,n}, \ldots, N_{K_n,n})$. If $I(\cdot)$ denotes the indicator function, then

$$\tau_1 = \sum_{i=1}^{K_n} I(N_{i,n} = 1) I(N_{i,N} = 1),$$

namely the number of sample uniques that are also population uniques (Bethlehem et al. [2], Skinner et al. [24]). Bayesian nonparametric inference for τ_1 relies on the specification of a (nonparametric) prior distribution on the discrete distribution P, which in turn leads to a prior model for the microdata partition.

2.1. The Pitman–Yor process prior

We assume the Pitman–Yor process as a prior model for the unknown discrete distribution P. A simple and intuitive definition of the Pitman–Yor process follows from its stick-breaking construction (Pitman [16]). For $\alpha \in [0, 1)$ and $\theta > -\alpha$ let: i) $(V_i)_{i\geq 1}$ be independent random variables such that V_i is distributed as a Beta distribution with parameter $(1 - \alpha, \theta + i\alpha)$; ii) $(Z_j)_{j\geq 1}$ be random variables, independent of the V_i's, and independent and identically distributed as a non-atomic distribution ν on $\mathbb{Z}$. If we set $p_1 = V_1$ and $p_j = V_j \prod_{1\leq i\leq j-1}(1 - V_i)$ for $j \geq 2$, which ensures that $\sum_{j\geq 1} p_j = 1$ almost surely, then $P_{\alpha,\theta} = \sum_{j\geq 1} p_j \delta_{Z_j}$ is a Pitman–Yor process on $\mathbb{Z}$ with "discount" α and scale θ. The Dirichlet process arises as a special case by letting $\alpha = 0$. The Pitman–Yor process generalizes the Dirichlet process by means of the "discount" α, which controls the tail behaviour of $P_{\alpha,\theta}$, ranging from geometric tails to heavy power-law tails. In particular, for $\alpha \in (0, 1)$, let $(p_{(j)})_{j\geq 1}$ be the random probabilities p_j's of $P_{\alpha,\theta}$ in decreasing order. Then, as $j \to +\infty$ the $p_{(j)}$'s follow a power-law distribution of exponent $\sigma = \alpha^{-1}$ (Pitman and Yor [18]). This shows that $\alpha \in (0, 1)$ tunes the power-law tail behaviour of $P_{\alpha,\theta}$ through small probabilities $p_{(j)}$'s: the larger α the heavier the tail of $P_{\alpha,\theta}$, whereas a geometric tail arises as $\alpha \to 0$.

According to de Finetti's representation theorem, a random sample from $P_{\alpha,\theta}$ is part of an exchangeable sequence of $\mathbb{Z}$-valued random variables $(X_i)_{i\geq 1}$ whose directing measure Π is the law of $P_{\alpha,\theta}$. Let $(X_1, \ldots, X_n)$ be a random sample from $P_{\alpha,\theta}$, i.e.

$$\begin{aligned} X_i \mid P_{\alpha,\theta} &\overset{\text{iid}}{\sim} P_{\alpha,\theta} \qquad i = 1, \ldots, n, \\ P_{\alpha,\theta} &\sim \Pi. \end{aligned} \tag{1}$$

Because of the discreteness of $P_{\alpha,\theta}$, the sample $(X_1,\ldots,X_n)$ induces a random partition of $\{1,\ldots,n\}$ into $K_n \leq n$ blocks, labelled by distinct symbols $\{Z_1^*,\ldots,Z_{K_n}^*\}$, with frequencies $(N_{1,n},\ldots,N_{K_n,n}) = (n_1,\ldots,n_k)$ such that $N_{i,n} \geq 1$ for $i = 1,\ldots,K_n$ and $\sum_{1\leq i\leq K_n} N_{i,n} = n$ (Pitman [Chapter 3, 17]) for a detailed account. A generative model for the X_i's, and hence for the induced random partition, is provided by the predictive distribution of the Pitman–Yor process, namely

$$\mathbb{P}(X_{n+1} \in \cdot \mid X_1,\ldots,X_n) = \frac{\theta + k\alpha}{\theta + n}\nu(\cdot) + \frac{1}{\theta + n}\sum_{i=1}^{k}(n_i - \alpha)\delta_{Z_j^*}(\cdot), \qquad (2)$$

for $n \geq 1$. That is, X_{n+1} is of a new symbol (block), namely a symbol not observed in the set $\{Z_1^*,\ldots,Z_{K_n}^*\}$, with probability $(\theta + k\alpha)/(\theta + n)$, or X_{n+1} is of symbol (block) Z_i^* with probability $(n_i - \alpha)/(\theta + n)$, for $i = 1,\ldots,k$. See Pitman [Chapter 3, 17] for a detailed account on the predictive distribution (2).

The predictive distribution of the Pitman–Yor process highlights the role of the "discount" parameter α in the sampling process: it drives a combined effect in terms of a reinforcement mechanism and the increase in the rate of generating new symbols. In particular, a new symbol z^* entering in the sample produces two effects: i) it is assigned a mass proportional to $(1-\alpha)$ to the z^*'s empirical component of (2); ii) it is assigned a mass proportional to α to the probability of generating new symbols in (2). That is, the probability mass assigned to the symbol z^*'s is less than proportional to 1, and the remaining probability mass is assigned to the probability of generating new symbols. The first effect gives rise to a reinforcement mechanism: the sampling procedure allocates more mass on symbols with higher frequencies. The second effect implies that the probability of generating new symbols, which overall still decreases as a function of n, is increased by $\alpha/(\theta+n+1)$. The larger α the stronger the reinforcement mechanism and the higher is the probability of new symbols. For $\alpha = 0$, that is under the Dirichlet process prior, everything is proportional to symbols' frequencies, which do not alter the probability of discovering new symbols. We refer to Bacallado et al. [1] for a detailed account on the predictive distribution (2), as well a generalizations thereof, and for characterizations of (2) with respect to the use of the sampling information, i.e. "sufficientness postulate", and of Pólya like urn schemes.

Remark 1. The power-law tail behaviour of the Pitman–Yor process emerges from the large n asymptotic behaviour of the number K_n of distinct symbols and the number $M_{r,n}$ of distinct symbols with frequency $r \geq 1$ in n random samples from $P_{\alpha,\theta}$. From Pitman [17, Theorem 3.8], K_n behaves as n^α for large n; this is the behaviour of the number of distinct symbols in n random samples from a power-law distribution of exponent $\sigma = \alpha^{-1}$. Moreover, from Pitman [17, Lemma 3.11] it holds that the proportion $M_{r,n}/K_n$ of distinct symbols with frequency r behaves as $r^{-\alpha-1}$ for large n and large r; this is, up to a constant or proportionality, the distribution of the number of distinct symbols

with frequency r in n random samples from a power-law distribution of exponent $\sigma = \alpha^{-1}$.

2.2. Posterior inference for τ_1

We consider microdata units to be modeled under the Bayesian nonparametric framework (1). That is, a population of $N \geq 1$ of microdata units is assumed to be a random sample $(X_1, \ldots, X_N)$ from a Pitman–Yor process, of which the first $n < N$ elements $(X_1, \ldots, X_n)$ are observable. We characterize the posterior distribution of τ_1, given $(X_1, \ldots, X_n)$. To introduce our main result, it is useful to recall the generalized factorial distribution (Charalambides [7, Chapter 2]). For a real a and $r \in \mathbb{N}$ let $(a)_{(r)}$ be the rising factorial, that is $(a)_{(0)} = 1$ and $(a)_{(r)} = \prod_{0 \leq i \leq r-1}(a+i)$ for $r \in \mathbb{N} \setminus \{0\}$, and for $a > 0$ and $r, s \in \mathbb{N}$ with $r \leq s$ let $\mathscr{C}(r, s; a)$ be the generalized factorial coefficient (Charalambides [7]), that is $\mathscr{C}(r, s; a) = \sum_{0 \leq i \leq s}(-1)^i \{i!(s-i)!\}^{-1}(-ia)_{(r)}$. For $r \in \mathbb{N}$, $b \in [0, 1]$ and $c > 0$, a random variable $U_{b,c,r}$ on $\{1, \ldots, r\}$ has a generalized factorial distribution if, for $x \in \{1, \ldots, r\}$

$$\mathbb{P}(U_{b,c,r} = x) = \frac{1}{(bc)_{(r)}} \mathscr{C}(r, x; b)(c)_{(x)}. \tag{3}$$

The next theorem provides the posterior distribution of τ_1, given $(X_1, \ldots, X_n)$, as a mixture of a (general) hypergeometric distribution (Johnson et al. [12, Chapter 6.2.5]) with respect to the generalized factorial distribution displayed in (3). Then, a Bayesian nonparametric estimator of τ_1 is given as the posterior mean.

Theorem 1. *For $N \geq 1$ let $(X_1, \ldots, X_N)$ be a random sample from $P_{\alpha,\theta}$, of which the first $n < N$ elements $(X_1, \ldots, X_n)$ are observable and featuring $M_{1,n} = m_1$ distinct symbols with frequency 1 (sample uniques). Then, for $x \in \{0, 1, \ldots, m_1\}$*

$$\mathbb{P}(\tau_1 = x \,|\, X_1, \ldots, X_n) = \sum_{u=1}^{N-n} \frac{\binom{\frac{\theta+n}{1-\alpha}-1}{x}\binom{u}{m_1-x}}{\binom{\frac{\theta+n}{1-\alpha}-1+u}{m_1}} \mathbb{P}(U_{1-\alpha,\frac{\theta+n}{1-\alpha},N-n} = u), \tag{4}$$

and

$$\hat{\tau}_1 = \mathbb{E}(\tau_1 \,|\, X_1, \ldots, X_n) = m_1 \frac{(\theta+\alpha+n-1)_{(N-n)}}{(\theta+n)_{(N-n)}}. \tag{5}$$

See Appendix A for the proof of Theorem 1. Theorem 1 is the first example in the literature to provide a closed-form posterior distribution of τ_1. This is critical to quantify, by means of Monte Carlo sampling, uncertainty of our Bayesian procedure through credible intervals; see Section 2.3 below. According to (4), for any fixed (α, θ), the number $M_{1,n} = m_1$ of sample uniques is sufficient for estimating τ_1. The estimator (5) is easy to implement, computationally efficient,

and scalable to massive datasets. Moreover, it has a simple interpretation as the proportion

$$w_{n,N}(\alpha,\theta)=\frac{(\theta+n-1+\alpha)_{(N-n)}}{(\theta+n)_{(N-n)}}\in(0,1),$$

of the number m_1 of sample uniques. The estimator (5) is somehow reminiscent of the "naive" nonparametric estimator (Bethlehem et al. [2], Skinner and Elliot [23]) of τ_1, namely

$$\bar{\tau}_1=m_1\frac{n}{N}.$$

In particular, $\hat{\tau}_1$ is a smoothed version of $\bar{\tau}_1$, where the smoothing acts by replacing the purely empirical proportion n/N with the parametric proportion $w_{n,N}(\alpha,\theta)$. For any fixed θ, n and N, the proportion $w_{n,N}(\alpha,\theta)$ increases in α, meaning that the larger α the higher $\hat{\tau}_1$. This behaviour, which agrees with the role of α discussed in Section 2.1, shows the effectiveness of the "discount" α in tuning the inference to the tail behaviour of the empirical distribution of microdata.

Remark 2. For $\alpha=0$, namely under the Dirichlet process prior, Theorem 1 simplifies remarkably. In particular, the posterior distribution (4) reduces to a (general) hypergeometric distribution. That is, by setting $\alpha=0$, Equation (4) reduces to

$$\mathbb{P}(\tau_1=x\,|\,X_1,\ldots,X_n)=\frac{\binom{\theta+n-1}{x}\binom{N-n}{m_1-x}}{\binom{\theta+N-1}{m_1}}. \tag{6}$$

for $x\in\{0,1,\ldots,m_1\}$. Moreover, by setting $\alpha=0$, Equation (5) reduces to the estimator of Samuels [21], namely $\hat{\tau}_1=m_1(\theta+n-1)/(\theta+N-1)$. Equation (4) thus completes the work of Samuels [21], where only the estimator $\hat{\tau}_1$ was provided.

By assuming both the sample and population to be large, it emerges: i) the critical influence of the "discount" α in estimating τ_1, with respect to the scale θ; ii) the crucial limitation of the estimator proposed in Samuels [21]. In particular, let $f\approx g$ meaning $f/g\to 1$. As $n,N\to+\infty$ with $n<N$, for any $x\in\{0,1,\ldots,m_1\}$

$$\mathbb{P}(\tau_1=x\,|\,X_1,\ldots,X_n)\approx\binom{m_1}{x}\left\{\left(\frac{n}{N}\right)^{1-\alpha}\right\}^{x}\left\{1-\left(\frac{n}{N}\right)^{1-\alpha}\right\}^{m_1-x}, \tag{7}$$

and hence

$$\hat{\tau}_1\approx m_1\left(\frac{n}{N}\right)^{1-\alpha}. \tag{8}$$

That is, for large n and N with $n<N$, the posterior distribution (4) admits a first order (local) approximation in terms of a Binomial distribution with parameters $\{m_1,(n/N)^{1-\alpha}\}$. See Appendix B for the proof of (7). This result shows that, in realistic scenarios, the "discount" α is the sole tuning parameter of our Bayesian nonparametric model. In other terms, for $\alpha=0$, namely under the Dirichlet process prior, the approximated estimator (8) reduces to the "naive"

estimator $\bar{\tau}_1$. Equivalently, for large n and N, the "naive" estimator $\bar{\tau}_1$ approximates the estimator of Samuels [21]. Therefore, in realistic scenarios, Samuel's estimator is a purely empirical estimator, meaning that no tuning parameters are available.

2.3. Computations

For any fixed $\alpha \in (0,1)$ and $\theta > -\alpha$, the estimator (5) can be easily evaluated for arbitrary values of n and N. Instead, the evaluation of the posterior distribution (4) might be numerically unstable for large n and N, due to the overwhelming computational burden for evaluating generalized factorial coefficients. To address this issue, we rely on Monte Carlo sampling of the posterior distribution (4) to obtain credible intervals for the estimator (5). By the mixture representation of (4), Monte Carlo sampling requires to sample from a (general) hypergeometric distribution and from a generalized factorial distribution. The former is straightforward, for arbitrary values of n and N, and routines are available in standard software. The latter becomes easy upon noticing that it coincides with the distribution of the number K_{N-n} of distinct symbols in $N-n$ random samples from a Pitman–Yor process with "discount" $(1-\alpha)$ and scale $(\theta+n)$. See Appendix C for a detailed explanation. For arbitrary values of n and N, Monte Carlo sampling of the distribution of K_{N-n} is straightforward by Algorithm 1, which exploits the predictive distribution (2) of the Pitman–Yor process.

Set $k = 1$;
for $i = 1$ *to* $N - n - 1$ **do**
  Sample a binary variable s with probability $\{\theta + n + (1-\alpha)k\}/(\theta + n + i)$;
  Set $k \leftarrow k + s$;
end
Return k.

Algorithm 1: Monte Carlo sampling of the mixing generalized factorial distribution.

To implement Theorem 1 we must specify the prior's parameters (α, θ), whose choice is critical for a correct estimation of τ_1. Two common approaches for estimating (α, θ) are: i) the hierarchical Bayes approach, which relies on Bayesian estimates obtained from the posterior distribution of (α, θ) with respect to suitable prior specification; ii) the empirical Bayes approach, which relies on estimates obtained by maximizing, with respect to (α, θ), the marginal likelihood of the observable sample. Here, we adopt the empirical Bayes approach. Let $(X_1, \ldots, X_n)$ feature $K_n = k$ distinct symbols with frequencies $(N_{1,n}, \ldots, N_{K_n,n}) = (n_1, \ldots, n_k)$. Pitman [16, Proposition 9] provides the likelihood function of $(X_1, \ldots, X_n)$, and the empirical Bayes approach reduces to

TABLE 1

Estimates of τ_1 for synthetic data. The parameters are σ (Zipf data) and π (Geometric data). PB-*1 is parametric Bayes of Bethlehem et al. [2];* PB-*2 is parametric Bayes of Skinner et al. [24], and* NEB *is nonparametric empirical Bayes of Camerlenghi et al. [6].*

DATA	m_1	τ_1	PITMAN-YOR	DIRICHLET PR.	PB-1	PB-2	NEB
SCENARIO I: $N = 10^6, n = 10^5$							
Zipf 1.25	10818	6914	6818 [6689, 6947]	1123 [1042, 1203]	1543	946	8328
Zipf 1.50	2045	941	948 [890, 1006]	206 [171, 241]	224	194	1403
Zipf 1.75	557	205	203 [174, 232]	56 [38, 75]	58	66	283
Zipf 2.00	230	80	74 [56, 93]	23 [12, 35]	22	30	198
Geom. 10^{-4}	9938	1027	1113 [1034, 1195]	1113 [1034, 1195]	4666	2095	740
Geom. 10^{-3}	949	91	96 [73, 120]	96 [73, 120]	335	167	67
SCENARIO II: $N = 5000, n = 500$							
Zipf 1.25	139	76	82 [67, 96]	16 [7, 27]	34	20	120
Zipf 1.50	62	23	28 [18, 38]	7 [1, 13]	12	7	51
Zipf 1.75	28	7	10 [4, 17]	3 [0, 8]	5	3	22
Zipf 2.00	11	3	3 [0, 7]	1 [0, 4]	2	1	6
Geom. 10^{-4}	482	391	365 [341, 388]	365 [341, 388]	181	196	467
Geom. 10^{-3}	387	95	129 [106, 153]	129 [106, 153]	160	158	320

solve:

$$(\hat{\alpha}, \hat{\theta}) = \arg\max_{(\alpha,\theta)} \left\{ \frac{\prod_{i=0}^{k-1}(\theta + i\alpha)}{(\theta)_{(n)}} \prod_{i=1}^{k} (1-\alpha)_{(n_i-1)} \right\}. \tag{9}$$

The optimization problem (9) can be solved numerically and efficiently even for large values of n, by means of routines available in standard softwares. We refer to Favaro and Naulet [9] for provable guarantees of the estimator $\hat{\alpha}$. Alternatively, one could specify a prior distribution on (α, θ). However, we found no relevant differences between the fully Bayes and the empirical Bayes approach, given that the posterior distribution of (α, θ) is highly concentrated, when n is large.

3. Illustrations

3.1. Simulated data

We consider synthetic data from two super-populations P. For the first super-population, we let the "true" probability masses $(p_j)_{j\geq 1}$ to be those of a Zipf distribution with index $\sigma > 1$, so that data are generated from the discrete distribution $P = \zeta(\sigma)^{-1} \sum_{j\geq 1} j^{-\sigma} \delta_{z_j}$, with $\zeta(\sigma) = \sum_{j\geq 1} j^{-\sigma}$. As we discussed in Section 2, this is the scenario in which a Pitman–Yor specification is recommended. We considered different values of $\sigma = 1.25, 1.50, 1.75, 2$, and different combinations of n and N. The prior's parameter (α, θ) is estimated through maximum likelihood; see Section 2.3. Table 1 reports estimates of τ_1, together with 99% credible intervals (within brackets), and the "true" value of τ_1. Credible intervals are obtained via Monte Carlo sampling of the posterior distribution

TABLE 2
Maximum likelihood estimate for the parameter (α, θ) of the Pitman–Yor model.

Param.	Zipf 1.25	Zipf 1.50	Zipf 1.75	Zipf 2.00	Geom. 10^{-4}	Geom. 10^{-4}
SCENARIO I: $N = 10^6, n = 10^5$						
$\hat{\alpha}$	0.80	0.67	0.56	0.51	0	0
$\hat{\theta}$	1.48	0.82	0.70	0.34	13559.80	1141.16
SCENARIO II: $N = 5000, n = 500$						
$\hat{\alpha}$	0.77	0.66	0.57	0.39	0	0
$\hat{\theta}$	1.89	0.98	0.52	0.90	13529.12	1753.06

(1), by means of the scheme described in Section 2.3. Table 2 reports the corresponding estimates of (α, θ) for the Pitman–Yor model. In all these scenarios, the Bayesian nonparametric estimator (5) is much closer to the "true" value of τ_1, compared to its partition-based competitors. In particular, the approaches of Bethlehem et al. [2], Skinner et al. [24] and Samuels [21] underestimate the "true" τ_1, whereas the approach of Camerlenghi et al. [6] tends to overestimate it.

For the second super-population, we let $P = \sum_{j \geq 1} \pi(1-\pi)^{j-1}\delta_{z_j}$, corresponding to a geometric distribution with parameter $\pi \in (0, 1)$. We consider two different values of $\pi = 10^{-3}, 10^{-4}$ and the same sample size n and a population size N as before. As we discussed in Section 2, this is the ideal setting for the Dirichlet process and this is indeed confirmed by Table 1. Moreover, the Pitman–Yor estimator reduces to the Dirichlet process since we obtain $\hat{\alpha} = 0$, as reported in Table 2.

3.2. The 2018 American Community Survey

We consider real data from the 2018 American Community Survey (Manrique-Vallier and Reiter [13], Carota et al. [4]). This dataset is a random sample of the American population (`usa.ipums.org/usa`). We regard the 2018 American Community data as a "population" of size $N = 2,432,323$, and we consider observable samples which are the 5% and 10% fractions of the population obtained by sampling at random $n = 121,616$ and $n = 243,232$ individual, respectively. We restricted the population to individuals older than 20, and we cross-classified the records according to the following variables: census region (9 levels), race (139 levels), and primary occupation (531 levels), obtaining $K_N = 60,215$ non empty classes.

As detailed in Section 2, the Pitman–Yor specification should be employed whenever the data follow a power-law behaviour. However, in real data problems such an assumption must be empirically validated. A simple approach is comparing the observed number m_r of distinct types with frequency $r = 1, \ldots, n$ against the model-based expected frequencies under a Pitman–Yor specification,

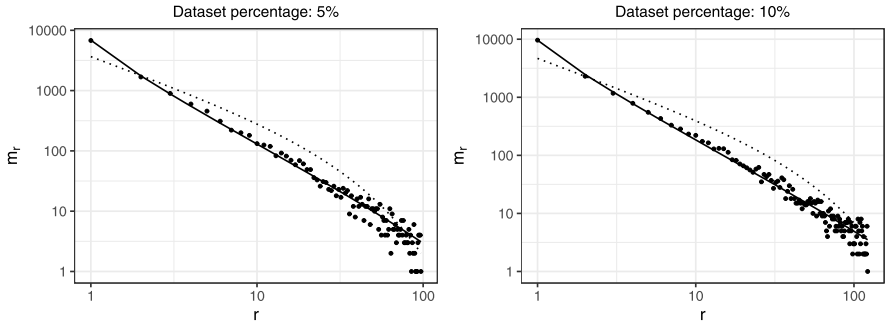


FIG 2. *Graphical representation in logarithmic scale of the number of distinct types m_r with frequency r (dots) and their expectations $\mathbb{E}(M_{r,n})$ under Pitman–Yor (solid line) and Dirichlet process (dotted line) models, relative to the 5% and 10% sample data from the American Community Survey.*

namely

$$\mathbb{E}(M_{r,n}) = \frac{\theta}{(\theta)_{(n)}}\binom{n}{r}(1-\alpha)_{(r-1)}(\theta+\alpha)_{(n-r)}, \qquad r = 1, \ldots, n,$$

where the parameters in the above formula are replaced by their maximum likelihood estimates; see also Favaro et al. [8] for further details. Poor in-sample fit strongly suggests that the corresponding disclosure risk assessment will be unreliable.

The observed values m_r for $r = 1, \ldots, n$ and their model-based estimates for the 5% and 10% fractions of the data from the American Community Survey presented in Section 2.2 are reported in Figure 2, both under a Pitman–Yor and Dirichlet process specification. These results confirm a very good in-sample fit for the Pitman–Yor. Conversely, the Dirichlet process seems unsuitable for this specific datasets. The prior's parameters α and θ are estimated through maximum likelihood; see Section 2.3. Results in Table 3 confirm what we observed for synthetic data, and in particular it is confirmed the superior empirical performance of our estimators, with respect to partition-based competitors. The approaches of Bethlehem et al. [2], Skinner et al. [24] and Samuels [21] underestimate the true τ_1, whereas the approach of Camerlenghi et al. [6] overestimates it.

TABLE 3

Estimates of τ_1 for real data the 2018 American Community Survey. The estimate PB-*1 refers to the parametric Bayes of Bethlehem et al. [2],* PB-*2 is the parametric Bayes of Skinner et al. [24], and* NEB *is the nonparametric empirical Bayes of Camerlenghi et al. [6].*

Data percentage	m_1	τ_1	Pitman–Yor	Dirichlet process	PB-1	PB-2	NEB
5%	6776	1447	1458 [1372, 1546]	349 [303, 397]	1427	425	3492
10%	9620	2852	2958 [2842, 3075]	979 [903, 1056]	1799	1059	4526

4. Discussion

In this paper, we considered the problem of Bayesian nonparametric estimation of τ_1, which is arguably the most popular measure of disclosure risk. Our study is motivated by an early work of Samuels [21], where empirical analyses showed that the use of Dirichlet process priors lead to underestimate τ_1 in many realistic scenario, with the underestimation getting worse as the tail behaviour of the empirical distribution of microdata gets heavier. Here, to overcome such an underestimation phenomenon, we proposed the use of the Pitman–Yor process prior, which generalizes the Dirichlet process prior through an additional "discount" parameter that allows to control the tail behaviour of the prior, ranging from geometric tails to heavy power-law tails. Under the Pitman–Yor process prior, we obtained a simple characterization of the posterior distribution of τ_1, in terms of a compound (general) hypergeometric distribution, and made use of the posterior mean as an estimator of τ_1. Such a novel estimator has all the desirable features as Samuels' estimator, including ease of implementation, computational efficiency and scalability to massive data, and, in addition, it allows to reduce its underestimation of τ_1 by tuning the "discount" parameter with respect to observable microdata. We presented an empirical analysis of our Bayesian nonparametric approach through synthetic data and real data, showing its effectiveness in reducing underestimation phenomenon of Samuels' approach.

While τ_1 is known to be the most popular measure of disclosure risk (Bethlehem et al. [2] and Skinner et al. [24]), one might consider alternative measures by broadening the definition of "uniqueness". For instance, Fienberg and Makov [11] considered a generalization of τ_1 which is defined in terms of the number of cells with frequency less or equal than 2. In general, one may consider the following measure

$$\tau_{p,q} = \sum_{i=1}^{K_n} I(N_{i,n} \leq p) I(N_{i,N} \leq p+q),$$

namely the number of cells with sample frequency less or equal than p which have population frequency less or equal than $p+q$. In particular, τ_1 corresponds to $\tau_{1,0}$. We refer to Appendix D for Bayesian nonparametric inference of $\tau_{1,q}$, which is arguably the most natural generalization of τ_1. It remains an open problem to adapt our Bayesian nonparametric approach to deal with structurally empty cells, i.e. structural zeros (Manrique-Vallier and Reiter [14]). In such a context, it may be useful to consider spike and slab generalizations of the Pitman-Yor process prior (Scarpa and Dunson [22], Canale et al. [3]). They consist in replacing the non-atomic distribution ν of the Pitman-Yor process prior with a distribution $\tilde{\nu}(\zeta) = \zeta\delta_0 + (1-\zeta)\nu$, with $\zeta \in [0,1]$ and ν being a non-atomic distribution. Then ζ may then be used to include the information on structural zeros, being interpretable as the proportion of structural zeros in the population.

Appendix A: Proof of Theorem 1

A.1. Generalized factorial coefficients

For $t \in \mathbb{R}$, $a > 0$ and $n \in \mathbb{N}_0$, let $(at)_{(n)}$ be the rising factorial of at of order n, i.e. $(at)_{(n)} = \prod_{0 \leq i \leq n-1}(at + i)$. The (n, k)-th generalized factorial coefficient, denoted by $\mathscr{C}(n, k; a)$, is the k-th coefficient in the expansion of $(at)_{(n)}$ into rising factorials, i.e.

$$(at)_{(n)} = \sum_{i=0}^{n} \mathscr{C}(n, i; a)(t)_{(i)}, \tag{10}$$

with $\mathscr{C}(0, 0; a) = 1$, $\mathscr{C}(n, 0; a) = 0$ for $n > 0$, $\mathscr{C}(n, i; a) = 0$ for $i > n$. For $b > 0$, let us consider the k-th coefficient in the expansion of $(at - b)_{(n)}$ into rising factorials, so that

$$(at - b)_{(n)} = \sum_{i=0}^{n} \mathscr{C}(n, i; a, b)(t)_{(i)}, \tag{11}$$

with $\mathscr{C}(0, 0; a, b) = 1$, $\mathscr{C}(n, 0; a, b) = (-b)_{(n)}$ for $n > 0$, $\mathscr{C}(n, i; a, b) = 0$ for $i > n$. The coefficient $\mathscr{C}(n, k; a, b)$ is referred to as the non-centered generalized factorial coefficient (Charalambides [7]). Here, it is useful to recall the following property

$$\mathscr{C}(n, i; b_1 b_2, b_1 r_2 + r_1) = \sum_{j=i}^{n} \mathscr{C}(n, j; b_1, r_1)\mathscr{C}(j, i; b_2, r_2), \tag{12}$$

for any $b_1, b_2 > 0$ and $r_1, r_2 > 0$. The convolutional identity (12) can be found in Charalambides [Chapter 2, 7] and plays a critical role in the proof of Theorem 1.

A.2. Generalized factorial and (general) hypergeometric distributions

The generalized factorial distribution (Charalambides [7, Chapter 2]) is defined by means of the identity (10), and it arises in the context of the classical coupon collector problem (Charalambides [7, Example 2.7]). For $r \in \mathbb{N}$ and $b, c > 0$, a random variable $U_{b,c,r}$ on the set $\{1, \ldots, r\}$ has a generalized factorial distribution if

$$\mathbb{P}(U_{b,c,r} = x) = \frac{1}{(bc)_{(r)}} \mathscr{C}(r, x; b)(c)_{(x)} I(x \in \{1, \ldots, r\}). \tag{13}$$

The (general) hypergeometric distribution (Johnson et al. [12, Chapter 6.2.5]) has the same form as the classical hypergeometric distribution, though with a more flexible parameterization. In particular, for $r, s \in \mathbb{N}$ and $a > 0$ such that $a > r$, a random variable $H_{a,r,s}$ on the set $\{0, 1, \ldots, r\}$ has a generalized factorial distribution if

$$\mathbb{P}(H_{a,r,s} = x) = \frac{\binom{a}{x}\binom{s}{r-x}}{\binom{a+s}{r}} I(x \in \{0, 1, \ldots, r\}). \tag{14}$$

Distributional properties, and moments, of the general hypergeometric distribution can be easily obtained from (14) (Johnson et al. [12, Chapter 6.3]). For $r, s \in \mathbb{N}$ with $s \leq r$ let $S(r,s)$ be the Stirling number of the second type (Charalambides [7, Chapter 2]), and let Γ denote the Gamma function. Then, for $z > 0$ it holds

$$\mathbb{E}\{(H_{a,r,s})^z\} = \sum_{i=1}^{z} S(z,i) i! \binom{r}{i} \frac{\Gamma(a+1+s-i)\,\Gamma(a+1)}{\Gamma(a+1-i)\,\Gamma(a+1+s)}. \tag{15}$$

We refer to Charalambides [7] and Johnson et al. [12] for a comprehensive account of the generalized factorial distribution and the (general) hypergeometric distribution.

A.3. Proof of Theorem 1

Let $(X_1, \ldots, X_n)$ be a random sample from the Pitman-Yor process $P_{\alpha,\theta}$, and let $(X_1, \ldots, X_n)$ feature $K_n = k$ distinct symbols, labelled by $\{Z_1^*, \ldots, Z_{K_n}^*\}$, with frequencies $\mathbf{N}_n = \mathbf{n}$, with $\mathbf{N}_n = (N_{1,n}, \ldots, N_{K_n,n})$, and $\mathbf{n} = (n_1, \ldots, n_k)$ be such that $N_{i,n} > 0$ and $\sum_{1 \leq i \leq K_n} N_{i,n} = n$. Moreover, for any $N > n$ let $(X_{n+1}, \ldots, X_N)$ be an additional random sample from $P_{\alpha,\theta}$, and let $N_{j,N-n} \geq 0$ be the number of records $X_{n+i}, i = 1 \ldots, N$ that coincide with the label $Z_j^*, j = 1, \ldots, K_n$. Moreover, let

$$V_{N-n} = N - n - \sum_{i=1}^{K_n} N_{i,N-n}$$

be the number of $X_{n+i}, i = 1, \ldots, N$ that do not coincide with any Z_j^*'s. To compute the posterior distribution of τ_1, we first determine its moment of order $z \geq 1$, i.e.,

$$\begin{aligned} &\mathbb{E}\{(\tau_1)^z \mid X_1, \ldots, X_n\} \\ &\quad = \mathbb{E}\{(\tau_1)^z \mid \mathbf{N}_n = \mathbf{n}, K_n = k\} \\ &\quad = \mathbb{E}\left\{ \left(\sum_{i=1}^{K_n} I(N_{i,n} = 1) I(N_{i,N-n} = 0) \right)^z \mid \mathbf{N}_n = \mathbf{n}, K_n = k \right\}. \end{aligned} \tag{16}$$

For $s, t \in \mathbb{N}$ with $s \leq t$ recall that $S(s,t)$ denotes the Stirling number of the second type (Charalambides [7, Chapter 2]), and let $\mathcal{C}_{t,s}$ denote a set of combination defined as follows: $\mathcal{C}_{t,0} = \emptyset$ and $\mathcal{C}_{t,s} = \{(c_1, \ldots, c_s) : c_i \in \{1, \ldots, t\}, c_i \neq c_j, \text{ if } i \neq j\}$ for any $s \geq 1$. Accordingly, Equation (16) admits the following expansion

$$\begin{aligned} &\mathbb{E}\{(\tau_1)^z \mid X_1, \ldots, X_n\} \\ &= \sum_{x=1}^{k} \sum_{i_1=1}^{z} \sum_{i_2=1}^{i_1-1} \cdots \sum_{i_{x-1}=1}^{i_{x-2}-1} \binom{r}{i_1} \binom{i_1}{i_2} \cdots \binom{i_{x-2}}{i_{x-1}} \end{aligned}$$

$$\times \sum_{(c_1,\ldots,c_x)\in\mathcal{C}_{k,x}} \mathbb{E}\left\{\prod_{t=1}^{x} I(N_{c_t,n}=1)I(N_{c_t,N-n}=0)^{i_{x-t}-i_{x-t+1}} \mid \mathbf{N}_n=\mathbf{n}, K_n=k\right\}$$

$$=\sum_{x=1}^{z} S(z,x)x! \sum_{(c_1,\ldots,c_x)\in\mathcal{C}_{k,x}} \mathbb{E}\left\{\prod_{t=1}^{x} I(N_{c_t,n}=1)I(N_{c_t,N-n}=0) \mid \mathbf{N}_n=\mathbf{n}, K_n=k\right\}$$

$$=\sum_{x=1}^{z} S(z,x)x! \sum_{(c_1,\ldots,c_x)\in\mathcal{C}_{k,x}} \prod_{t=1}^{x} I(N_{c_t,n}=1)\mathbb{E}\left\{\prod_{t=1}^{x} I(N_{c_t,N-n}=0) \mid \mathbf{N}_n=\mathbf{n}, K_n=k\right\}$$

$$=\sum_{x=1}^{z} S(z,x)x! \tag{17}$$

$$\times \sum_{(c_1,\ldots,c_x)\in\mathcal{C}_{k,x}} \prod_{t=1}^{x} I(N_{c_t,n}=1)\mathbb{P}(N_{c_1,N-n}=0,\ldots,N_{c_x,N-n}=0 \mid \mathbf{N}_n=\mathbf{n}, K_n=k).$$

The conditional probability (17) can be computed by a direct application of Favaro et al. [8, Lemma 1]. In particular, from Favaro et al. [8, Equation 38 and Equation 40]

$$\mathbb{P}(N_{c_1,N-n}=0,\ldots,N_{c_x,N-n}=0 \mid \mathbf{N}_n=\mathbf{n}, K_n=k, V_{N-n}=v) = \frac{(n-\sum_{i=1}^{x} n_{c_i}-(k-x)\alpha)_{(N-n-v)}}{(n-k\alpha)_{(N-n-v)}};$$

and

$$\mathbb{P}(V_{N-n}=v \mid \mathbf{N}_n=\mathbf{n}, K_n=k) = \binom{N-n}{v}(n-k\alpha)_{(N-n-v)} \sum_{j=0}^{v} \frac{\frac{\prod_{i=0}^{k+j-1}(\theta+i\alpha)}{(\theta)_{(n+(N-n))}}}{\frac{\prod_{i=0}^{k-1}(\theta+i\alpha)}{(\theta)_{(n)}}} \frac{\mathscr{C}(v,j;\alpha)}{\alpha^j}.$$

Then,

$$\begin{aligned}
&\mathbb{P}(N_{c_1,N-n}=0,\ldots,N_{c_x,N-n}=0 \mid \mathbf{N}_n=\mathbf{n}, K_n=k) \qquad (18)\\
&= \sum_{v=0}^{N-n} \binom{N-n}{v}(n-k\alpha)_{(N-n-v)} \sum_{j=0}^{v} \frac{\frac{\prod_{i=0}^{k+j-1}(\theta+i\alpha)}{(\theta)_{(n+(N-n))}}}{\frac{\prod_{i=0}^{k-1}(\theta+i\alpha)}{(\theta)_{(n)}}} \frac{\mathscr{C}(v,j;\alpha)}{\alpha^j}\\
&\qquad \times \frac{(n-\sum_{i=1}^{x} n_{c_i}-(k-x)\alpha)_{(N-n-v)}}{(n-k\alpha)_{(N-n-v)}}\\
&= \sum_{j=0}^{N-n} \frac{1}{\alpha^j} \frac{\frac{\prod_{i=0}^{k+j-1}(\theta+i\alpha)}{(\theta)_{(n+(N-n))}}}{\frac{\prod_{i=0}^{k-1}(\theta+i\alpha)}{(\theta)_{(n)}}} \sum_{v=j}^{N-n} \binom{N-n}{s}(n-k\alpha)_{(N-n-v)}\mathscr{C}(v,j;\alpha)\\
&\qquad \times \frac{(n-\sum_{i=1}^{x} n_{c_i}-(k-x)\alpha)_{(N-n-v)}}{(n-k\alpha)_{(N-n-v)}}
\end{aligned}$$

$$= \sum_{j=0}^{N-n} \frac{1}{\alpha^j} \frac{\frac{\prod_{i=0}^{k+j-1}(\theta+i\alpha)}{(\theta)_{(n+(N-n))}}}{\frac{\prod_{i=0}^{k-1}(\theta+i\alpha)}{(\theta)_{(n)}}} \sum_{v=j}^{N-n} \binom{N-n}{v} \mathscr{C}(v,j;\alpha)$$

$$\times (n - \sum_{i=1}^{x} n_{c_i} - (k-x)\alpha)_{(N-n-v)} \tag{19}$$

$$= \sum_{j=0}^{N-n} \frac{1}{\alpha^j} \frac{\frac{\prod_{i=0}^{k+j-1}(\theta+i\alpha)}{(\theta)_{(n+(N-n))}}}{\frac{\prod_{i=0}^{k-1}(\theta+i\alpha)}{(\theta)_{(n)}}} \sum_{v=j}^{N-n} \binom{N-n}{v} \mathscr{C}(v,j;\alpha)$$

$$\times \mathscr{C}(N-n-v, 0, \alpha, -n + \sum_{i=1}^{x} n_{c_i} + (k-x)\alpha). \tag{20}$$

Then, by the application of the convolutional identity (12) to the sum over v, we get

$$\begin{aligned}
&\mathbb{P}(N_{c_1,N-n} = 0, \ldots, N_{c_x,N-n} = 0 \mid \mathbf{N}_n = \mathbf{n}, K_n = k) \\
&\quad = \sum_{j=0}^{N-n} \frac{\frac{\prod_{i=0}^{k+j-1}(\theta+i\alpha)}{(\theta)_{(n+(N-n))}}}{\frac{\prod_{i=0}^{k-1}(\theta+i\alpha)}{(\theta)_{(n)}}} \frac{\mathscr{C}(N-n, j; \alpha, -n + \sum_{i=1}^{x} n_{c_i} + (k-x)\alpha)}{\alpha^j} \\
&\quad = \frac{1}{(\theta+n)_{(N-n)}} \sum_{j=0}^{N-n} \left(\frac{\theta}{\alpha} + k\right)_{(j)} \mathscr{C}(N-n, j; \alpha, -n + \sum_{i=1}^{x} n_{c_i} + (k-x)\alpha).
\end{aligned}$$

Therefore, by the application of the identity (11) to the sum over j, we obtain that

$$\begin{aligned}
&\mathbb{P}(N_{c_1,N-n} = 0, \ldots N_{c_x,N-n} = 0 \mid \mathbf{N}_n = \mathbf{n}, K_n = k) \\
&\qquad = \frac{(\theta + n - \sum_{i=1}^{x} n_{c_i} + x\alpha)_{(N-n)}}{(\theta+n)_{(N-n)}}.
\end{aligned}$$

By a direct combination of Equation (17) and Equation (18), the moment formula (16) is

$$\begin{aligned}
&\mathbb{E}\{(\tau_1)^z \mid X_1, \ldots, X_n\} \\
&\quad = \sum_{x=1}^{z} S(z,x) x! \sum_{(c_1,\ldots,c_x) \in \mathcal{C}_{k,x}} \prod_{t=1}^{x} I(N_{c_t,n} = 1) \\
&\qquad \times \mathbb{P}(N_{c_1,N-n} = 0, \ldots, N_{c_x,N-n} = 0 \mid \mathbf{N}_n = \mathbf{n}, K_n = k) \\
&\quad = \sum_{x=1}^{z} S(z,x) x! \sum_{(c_1,\ldots,c_x) \in \mathcal{C}_{k,x}} \prod_{t=1}^{x} I(N_{c_t,n} = 1) \frac{(\theta + n - \sum_{i=1}^{x} n_{c_i} + x\alpha)_{(N-n)}}{(\theta+n)_{(N-n)}}
\end{aligned}$$

$$= \sum_{x=1}^{z} S(z,x)x!\binom{m_1}{x}\frac{(\theta+n-x+x\alpha)_{(N-n)}}{(\theta+n)_{(N-n)}}$$
$$= \frac{1}{(\theta+n)_{(N-n)}}\sum_{x=1}^{z} S(z,x)x!\binom{m_1}{x}\left\{(1-\alpha)\left(\frac{\theta+n}{1-\alpha}-x\right)\right\}_{(N-n)},$$

and from (10)

$$\begin{aligned}
&\mathbb{E}\{(\tau_1)^z \mid X_1,\ldots,X_n\}\\
&\quad= \frac{1}{(\theta+n)_{(N-n)}}\sum_{x=1}^{z} S(z,x)x!\binom{m_1}{x}\sum_{i=1}^{N-n}\mathscr{C}(n,i;1-\alpha)\left(\frac{\theta+n}{1-\alpha}-x\right)_{(i)}\\
&\quad= \frac{1}{(\theta+n)_{(N-n)}}\sum_{i=1}^{N-n}\mathscr{C}(n,i;1-\alpha)\sum_{x=1}^{z} S(z,x)x!\binom{m_1}{x}\frac{\Gamma\left(\frac{\theta+n}{1-\alpha}+i-x\right)}{\Gamma\left(\frac{\theta+n}{1-\alpha}-x\right)}\\
&\quad= \frac{1}{(\theta+n)_{(N-n)}}\sum_{i=1}^{N-n}\mathscr{C}(n,i;1-\alpha)\frac{\Gamma\left(\frac{\theta+n}{1-\alpha}+i\right)}{\Gamma\left(\frac{\theta+n}{1-\alpha}\right)}\\
&\qquad\times\sum_{x=1}^{z} S(z,x)x!\binom{m_1}{x}\frac{\Gamma\left(\frac{\theta+n}{1-\alpha}+i-x\right)\Gamma\left(\frac{\theta+n}{1-\alpha}\right)}{\Gamma\left(\frac{\theta+n}{1-\alpha}-x\right)\Gamma\left(\frac{\theta+n}{1-\alpha}+i\right)}
\end{aligned}$$

[by the application of (15)]

$$= \frac{1}{(\theta+n)_{(N-n)}}\sum_{i=1}^{N-n}\mathscr{C}(n,i;1-\alpha)\left(\frac{\theta+n}{1-\alpha}\right)_{(i)}\mathbb{E}\{(H_{\frac{\theta+n}{1-\alpha}-1,m_1,i})^z\} \tag{21}$$

[by the definition of generalized factorial distribution (13)]

$$= \sum_{i=1}^{N-n}\mathbb{E}\{(H_{\frac{\theta+n}{1-\alpha}-1,m_1,i})^z\}\mathbb{P}(U_{1-\alpha,\frac{\theta+n}{1-\alpha},N-n}=i). \tag{22}$$

According to the above expression for $\mathbb{E}\{(\tau_1)^z \mid X_1,\ldots,X_n\}$, the proof of Theorem 1 is completed by using the definition of (general) hypergeometric distribution (14).

Appendix B: Proofs of Equation (7) and Equation (8)

Let $(X_1,\ldots,X_n)$ be a random sample from $P_{\alpha,\theta}$, and let $(X_1,\ldots,X_n)$ feature $K_n=k$ distinct symbols with $\mathbf{N}_n=(N_{1,n},\ldots,N_{K_n,n})$ corresponding frequencies, $\mathbf{n}=(n_1,\ldots,n_k)$ such that $N_{i,n}>0$ and $\sum_{1\le i\le K_n}N_{i,n}=n$. From the proof on Theorem 1,

$$\mathbb{E}\{(\tau_1)^z \mid X_1,\ldots,X_n\} = \sum_{i=1}^{z} S(z,x)i!\binom{m_1}{i}\frac{(\theta+n-i+i\alpha)_{(N-n)}}{(\theta+n)_{(N-n)}}. \tag{23}$$

Recall that by means of Stirling formula $\Gamma(n+i)/\Gamma(n) \approx n^i$ as $n \to +\infty$ is a first order approximation of the Gamma function. By applying it to (23), as $n \to +\infty$ and $N \to +\infty$.

$$\begin{aligned}
&\mathbb{E}\{(\tau_1)^z \mid X_1, \ldots, X_n\} \\
&= \sum_{i=1}^{z} S(z,i) i! \binom{m_1}{i} \frac{(\theta + n - i + i\alpha)_{(N-n)}}{(\theta+n)_{(N-n)}} \\
&= \sum_{i=1}^{z} S(z,i) i! \binom{m_1}{i} \frac{\frac{\Gamma(\theta+N-i+i\alpha)}{\Gamma(\theta+n-i+i\alpha)}}{\frac{\Gamma(\theta+N)}{\Gamma(\theta+n)}} \\
&\approx \sum_{i=1}^{z} S(z,i) i! \binom{m_1}{i} \left\{ \left(\frac{n}{N}\right)^{1-\alpha} \right\}^i . \qquad (24)
\end{aligned}$$

Equation (24) is the moment of order z of a Binomial random variable with parameter $(m_1, (n/N)^{1-\alpha})$, with m_1 being the number of trials and $(n/N)^{1-\alpha}$ being the probability of success in a trial. This completes the proof of Equation (7) and Equation (8).

Appendix C: On the distribution of $U_{1-\alpha,\frac{\theta+n}{1-\alpha},N-n}$

Let $(X_1, \ldots, X_n)$ be a random sample from $P_{\alpha,\theta}$, and let $(X_1, \ldots, X_n)$ feature $K_n = k$ distinct symbols with corresponding frequencies $\mathbf{N}_n = \mathbf{n}$, where $\mathbf{N}_n = (N_{1,n}, \ldots, N_{K_n,n})$ and $\mathbf{n} = (n_1, \ldots, n_k)$ such that $N_{i,n} > 0$ and $\sum_{1 \le i \le K_n} N_{i,n} = n$. The distribution of K_n is known from Pitman [17, Chapter 3]. In particular, for $x \in \{1, \ldots, n\}$

$$\mathbb{P}(K_n = x) = \frac{(\theta/\alpha)_{(x)}}{(\theta)_{(n)}} \mathscr{C}(n, x; \alpha). \qquad (25)$$

According to (25), the distribution of $U_{1-\alpha,\frac{\theta+n}{1-\alpha},N-n}$ coincides with the distribution of the number K_{N-n} distinct symbols in $N-n$ random samples from $P_{1-\alpha,\theta+n}$.

Appendix D: Bayesian nonparametric inference for $\tau_{1,q}$

Under the Pitman-Yor process prior, we characterize the posterior distribution of $\tau_{1,q}$ through its moments; this leads to a Bayesian nonparametric estimator of $\tau_{1,q}$ in terms of the posterior mean. The proof is along lines similar to the proof of Theorem 1. Let $(X_1, \ldots, X_n)$ be a random sample from the Pitman-Yor process $P_{\alpha,\theta}$, and let $(X_1, \ldots, X_n)$ feature $K_n = k$ distinct symbols, labelled by $\{Z_1^*, \ldots, Z_{K_n}^*\}$, with frequencies $\mathbf{N}_n = \mathbf{n}$, with $\mathbf{N}_n = (N_{1,n}, \ldots, N_{K_n,n})$, and $\mathbf{n} = (n_1, \ldots, n_k)$ be such that $N_{i,n} > 0$ and $\sum_{1 \le i \le K_n} N_{i,n} = n$. Moreover, for any $N > n$ let $(X_{n+1}, \ldots, X_N)$ be an additional random sample from $P_{\alpha,\theta}$, and let $N_{j,N-n} \ge 0$ be the number of records $X_{n+i}, i = 1 \ldots, N$ that coincide with

the label $Z_j^*, j = 1, \ldots, K_n$. Moreover, let $V_{N-n} = N - n - \sum_{1 \le i \le K_n} N_{i,N-n}$ be the number of $X_{n+i}, i = 1, \ldots, N$ that do not coincide with any Z_j^*'s. To compute the posterior distribution of $\tau_{1,q}$, we first determine its moment of order $z \ge 1$, i.e.,

$$\begin{aligned} &\mathbb{E}\{(\tau_{1,q})^z \mid X_1, \ldots, X_n\} \\ &\quad = \mathbb{E}\{(\tau_{1,q})^z \mid \mathbf{N}_n = \mathbf{n}, K_n = k\} \\ &\quad = \mathbb{E}\left\{ \left(\sum_{i=1}^{K_n} I(N_{i,n} = 1) I(N_{i,N-n} \le q) \right)^z \mid \mathbf{N}_n = \mathbf{n}, K_n = k \right\}. \end{aligned} \tag{26}$$

For $s, t \in \mathbb{N}$ with $s \le t$ recall that $S(s,t)$ denotes the Stirling number of the second type (Charalambides [7, Chapter 2]), and let $\mathcal{C}_{t,s}$ denote a set of combination defined as follows: $\mathcal{C}_{t,0} = \emptyset$ and $\mathcal{C}_{t,s} = \{(c_1, \ldots, c_s) : c_i \in \{1, \ldots, t\}, c_i \ne c_j, \text{ if } i \ne j\}$ for any $s \ge 1$. Accordingly, Equation (26) admits the following expansion

$$\begin{aligned} &\mathbb{E}\{(\tau_{1,q})^z \mid X_1, \ldots, X_n\} \\ &= \mathbb{E}\left\{ \left[\sum_{i=1}^{k} \left(\sum_{h=0}^{q} I(N_{i,n}=1) I(N_{i,N-n}=h) \right) \right]^z \mid \mathbf{N}_n = \mathbf{n}_n, K_n = k \right\} \\ &= \sum_{x=1}^{k} \sum_{i_1=1}^{z} \sum_{i_2=1}^{i_1-1} \cdots \sum_{i_{x-1}=1}^{i_{x-2}-1} \binom{z}{i_1} \binom{i_1}{i_2} \cdots \binom{i_{x-2}}{i_{x-1}} \\ &\quad \times \sum_{(c_1,\ldots,c_x) \in \mathcal{C}_{k,x}} \mathbb{E}\left\{ \prod_{t=1}^{x} \left(\sum_{h=0}^{q} I(N_{c_t,n}=1) I(N_{c_t,N-n}=h) \right)^{i_{x-t}-i_{x-t+1}} \mid \mathbf{N}_n = \mathbf{n}_n, K_n = k \right\} \\ &= \sum_{x=1}^{k} \sum_{i_1=1}^{z} \sum_{i_2=1}^{i_1-1} \cdots \sum_{i_{x-1}=1}^{i_{x-2}-1} \binom{z}{i_1} \binom{i_1}{i_2} \cdots \binom{i_{x-2}}{i_{x-1}} \\ &\quad \times \sum_{(c_1,\ldots,c_x) \in \mathcal{C}_{k,x}} \mathbb{E}\left\{ \prod_{t=1}^{x} \left(\sum_{h=0}^{q} I(N_{c_t,n}=1) I(N_{c_t,N-n}=h) \right) \mid \mathbf{N}_n = \mathbf{n}_n, K_n = k \right\} \\ &= \sum_{x=1}^{z} S(z,x) x! \\ &\quad \times \sum_{(c_1,\ldots,c_x) \in \mathcal{C}_{k,x}} \mathbb{E}\left\{ \prod_{t=1}^{x} \left(\sum_{h=0}^{q} I(N_{c_t,n}=1) I(N_{c_t,N-n}=h) \right) \mid \mathbf{N}_n = \mathbf{n}_n, K_n = k \right\}. \end{aligned}$$

Now, we define the (cartesian product) set $\mathcal{H}_{q,x} = \{0, \ldots, q\}^x$, such that we can write

$$\begin{aligned} &\mathbb{E}\{(\tau_{1,q})^z \mid X_1, \ldots, X_n\} \\ &= \sum_{x=1}^{z} S(z,x) x! \end{aligned}$$

$$\times \sum_{(c_1,\ldots,c_x)\in\mathcal{C}_{k,x}} \sum_{(h_1,\ldots,h_x)\in\mathcal{H}_{q,x}} \mathbb{E}\left\{\prod_{t=1}^{x}(I(N_{c_t,n}=1)I(N_{c_t,N-n}=h_t))|\mathbf{N}_n=\mathbf{n}_n, K_n=k\right\}$$

$$= \sum_{x=1}^{z} S(z,x)x! \sum_{(h_1,\ldots,h_x)\in\mathcal{H}_{q,x}}$$

$$\times \sum_{(c_1,\ldots,c_x)\in\mathcal{C}_{k,x}} \prod_{t=1}^{x} I(N_{c_t,n}=1)\mathbb{E}\left\{\prod_{t=1}^{x}(I(N_{c_t,N-n}=h_t)) \mid \mathbf{N}_n=\mathbf{n}_n, K_n=k\right\}$$

$$= \sum_{x=1}^{z} S(z,x)x! \sum_{(h_1,\ldots,h_x)\in\mathcal{H}_{q,x}} \tag{27}$$

$$\times \sum_{(c_1,\ldots,c_x)\in\mathcal{C}_{k,x}} \prod_{t=1}^{x} I(N_{c_t,n}=1)\mathbb{P}(N_{c_1,N-n}=h_1,\ldots,N_{c_x,N-n}=h_x|\mathbf{N}_n=\mathbf{n}_n, K_n=k).$$

The conditional probability in (27) can be computed from Lemma 1 in Favaro et al. (2013). In particular, from Equation 38 and Equation 40 in Favaro et al. (2013) we have

i)

$$\begin{aligned}&\Pr(N_{c_1,N-n}=h_1,\ldots,N_{c_x,N-n}=h_x \,|\, \mathbf{N}_n=\mathbf{n}_n, K_n=k, V_{N-n}=v)\\ &\quad= \frac{(N-n-v)!}{(N-n-v-\sum_{t=1}^{x}h_t)!}\prod_{t=1}^{x}\frac{(n_{c_t}-\alpha)_{(h_t)}}{h_t!}\\ &\qquad\times \frac{(n-\sum_{t=1}^{x}n_{c_t}-(k-x)\alpha)_{(N-n-v-\sum_{t=1}^{x}h_t)}}{(n-k\alpha)_{(N-n-v)}}\end{aligned}$$

ii)

$$\begin{aligned}&\Pr(V_{N-n}=v \,|\, \mathbf{N}_n=\mathbf{n}_n, K_n=k)\\ &\quad= \binom{N-n}{v}(n-k\alpha)_{(N-n-v)}\sum_{j=0}^{v}\frac{\frac{\prod_{i=0}^{k+j-1}(\theta+i\alpha)}{(\theta)_{(n+(N-n))}}}{\frac{\prod_{i=0}^{k-1}(\theta+i\alpha)}{(\theta)_{(n)}}}\frac{\mathscr{C}(v,j;\alpha)}{\alpha^j},\end{aligned}$$

and

$$\begin{aligned}&\Pr(N_{c_1,N-n}=h_1,\ldots,N_{c_x,N-n}=h_x \,|\, \mathbf{N}_n=\mathbf{n}_n, K_n=k) \qquad (28)\\ &\quad= \sum_{v=0}^{N-n}\binom{N-n}{v}(n-k\alpha)_{(N-n-v)}\sum_{j=0}^{v}\frac{\frac{\prod_{i=0}^{k+j-1}(\theta+i\alpha)}{(\theta)_{(n+(N-n))}}}{\frac{\prod_{i=0}^{k-1}(\theta+i\alpha)}{(\theta)_{(n)}}}\frac{\mathscr{C}(v,j;\alpha)}{\alpha^j}\\ &\qquad\times \frac{(N-n-v)!}{(N-n-v-\sum_{t=1}^{x}h_t)!}\prod_{t=1}^{x}\frac{(n_{c_t}-\alpha)_{(h_t)}}{h_t!}\\ &\qquad\quad\times \frac{(n-\sum_{t=1}^{x}n_{c_t}-(k-x)\alpha)_{(N-n-v-\sum_{t=1}^{x}h_t)}}{(n-k\alpha)_{(N-n-v)}}\end{aligned}$$

$$
\begin{aligned}
&= \sum_{j=0}^{N-n} \frac{1}{\alpha^j} \frac{\frac{\prod_{i=0}^{k+j-1}(\theta+i\alpha)}{(\theta)_{(n+(N-n))}}}{\frac{\prod_{i=0}^{k-1}(\theta+i\alpha)}{(\theta)_{(n)}}} \sum_{v=j}^{N-n} \binom{N-n}{v} (n-k\alpha)_{(N-n-v)} \mathscr{C}(v,j;\alpha) \\
&\quad \times \frac{(N-n-v)!}{(N-n-v-\sum_{t=1}^{x} h_t)!} \prod_{t=1}^{x} \frac{(n_{c_t}-\alpha)_{(h_t)}}{h_t!} \\
&\qquad \times \frac{(n-\sum_{t=1}^{x} n_{c_t}-(k-x)\alpha)_{(N-n-v-\sum_{t=1}^{x} h_t)}}{(n-k\alpha)_{(N-n-v)}} \\
&= \frac{(N-n)!}{(N-n-\sum_{t=1}^{x} h_t)!} \prod_{t=1}^{x} \frac{(n_{c_t}-\alpha)_{(h_t)}}{h_t!} \sum_{j=0}^{N-n} \frac{1}{\alpha^j} \frac{\frac{\prod_{i=0}^{k+j-1}(\theta+i\alpha)}{(\theta)_{(n+(N-n))}}}{\frac{\prod_{i=0}^{k-1}(\theta+i\alpha)}{(\theta)_{(n)}}} \\
&\quad \times \sum_{v=j}^{N-n} \mathscr{C}(v,j;\alpha) \frac{(N-n-\sum_{t=1}^{x} h_t)!}{(N-n-v-\sum_{t=1}^{x} h_t)!v!} \\
&\qquad \times (n-\sum_{t=1}^{x} n_{c_t}-(k-x)\alpha)_{(N-n-v-\sum_{t=1}^{x} h_t)} \\
&= \frac{(N-n)!}{(N-n-\sum_{t=1}^{x} h_t)!} \prod_{t=1}^{x} \frac{(n_{c_t}-\alpha)_{(h_t)}}{h_t!} \sum_{j=0}^{N-n} \frac{1}{\alpha^j} \frac{\frac{\prod_{i=0}^{k+j-1}(\theta+i\alpha)}{(\theta)_{(n+(N-n))}}}{\frac{\prod_{i=0}^{k-1}(\theta+i\alpha)}{(\theta)_{(n)}}} \\
&\quad \times \sum_{v=j}^{N-n} \binom{N-n-\sum_{t=1}^{x} h_t}{v} \mathscr{C}(v,j;\alpha) \\
&\qquad \times \mathscr{C}(N-n-v-\sum_{t=1}^{x} h_t, 0; \alpha, -n+\sum_{t=1}^{x} n_{c_t}+(k-x)\alpha) \\
&= \frac{(N-n)!}{(N-n-\sum_{t=1}^{x} h_t)!} \prod_{t=1}^{x} \frac{(n_{c_t}-\alpha)_{(h_t)}}{h_t!} \sum_{j=0}^{N-n} \frac{1}{\alpha^j} \frac{\frac{\prod_{i=0}^{k+j-1}(\theta+i\alpha)}{(\theta)_{(n+(N-n))}}}{\frac{\prod_{i=0}^{k-1}(\theta+i\alpha)}{(\theta)_{(n)}}} \\
&\quad \times \mathscr{C}(N-n-\sum_{t=1}^{x} h_t, j; \alpha, -n+\sum_{t=1}^{x} n_{c_t}+(k-x)\alpha) \\
&= \frac{(N-n)!}{(N-n-\sum_{t=1}^{x} h_t)!} \prod_{t=1}^{x} \frac{(n_{c_t}-\alpha)_{(h_t)}}{h_t!} \frac{1}{(\theta+n)_{(N-n)}} \\
&\quad \times \sum_{j=0}^{N-n} \left(\frac{\theta}{\alpha}+k\right)_{(j)} \mathscr{C}(N-n-\sum_{t=1}^{x} h_t, j; \alpha, -n+\sum_{t=1}^{x} n_{c_t}+(k-x)\alpha) \\
&= \frac{(N-n)!}{(N-n-\sum_{t=1}^{x} h_t)!} \prod_{t=1}^{x} \frac{(n_{c_t}-\alpha)_{(h_t)}}{h_t!} \frac{1}{(\theta+n)_{(N-n)}} \\
&\quad \times (\theta+n-\sum_{i=1}^{x} n_{c_t}+x\alpha)_{(N-n-\sum_{t=1}^{x} h_t)}.
\end{aligned}
$$

Then, by combining Equation (27) with Equation (28) we can write the following identities

$$
\begin{aligned}
&\mathbb{E}\{(\tau_{1,q})^z \,|\, X_1,\ldots,X_n\}\\
&=\sum_{x=1}^{z} S(z,x)x! \sum_{(h_1,\ldots,h_x)\in\mathcal{H}_{q,x}}\\
&\quad\times\sum_{(c_1,\ldots,c_x)\in\mathcal{C}_{k,x}} \prod_{t=1}^{x} I(N_{c_t,n}=1)\mathbb{P}(N_{c_1,N-n}=h_1,\ldots,N_{c_x,N-n}=h_x|\mathbf{N}_n=\mathbf{n}_n,K_n=k)\\
&=\sum_{x=1}^{z} S(z,x)x! \sum_{(h_1,\ldots,h_x)\in\mathcal{H}_{q,x}}\\
&\quad\times \sum_{(c_1,\ldots,c_x)\in\mathcal{C}_{k,x}} \prod_{t=1}^{x} I(N_{c_t,n}=1)\frac{(N-n)!}{(N-n-\sum_{t=1}^{x}h_t)!}\\
&\qquad\times \prod_{t=1}^{x}\frac{(n_{c_t}-\alpha)_{(h_t)}}{h_t!}\frac{(\theta+n-\sum_{i=1}^{x}n_{c_t}+x\alpha)_{(N-n-\sum_{t=1}^{x}h_t)}}{(\theta+n)_{(N-n)}}\\
&=\sum_{x=1}^{r} S(r,x)x!\binom{m_1}{x} \sum_{(h_1,\ldots,h_x)\in\mathcal{H}_{q,x}}\\
&\quad\times \frac{(N-n)!}{(N-n-\sum_{t=1}^{x}h_t)!}\prod_{t=1}^{x}\frac{(1-\alpha)_{(h_t)}}{h_t!}\frac{(\theta+n-x+x\alpha)_{(N-n-\sum_{t=1}^{x}h_t)}}{(\theta+n)_{(N-n)}}.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\mathbb{E}\{(\tau_{1,q})_{[z]} \,|\, X_1,\ldots,X_n\} \qquad (29)\\
&\quad=\mathbb{E}\left\{\left(\sum_{i=1}^{k}\left(\sum_{h=0}^{q} I(N_{i,n}=1)I(N_{i,N-n}=h)\right)\right)_{[z]} \,|\, \mathbf{N}_n=\mathbf{n}_n,K_n=k\right\}\\
&\quad=z!\binom{m_1}{z}\sum_{i_1=0}^{q}\cdots\sum_{i_z=0}^{q}\\
&\qquad\times\frac{(N-n)!}{(N-n-\sum_{t=1}^{z}i_t)!}\prod_{t=1}^{z}\frac{(1-\alpha)_{(i_t)}}{i_t!}\frac{(\theta+n-z+z\alpha)_{(N-n-\sum_{t=1}^{z}i_t)}}{(\theta+n)_{(N-n)}}.
\end{aligned}
$$

Equation (29) leads, by means of standard arguments on inversion formula, to the calculation of the conditional distribution of $\tau_{1,q}$ given $(X_1,\ldots,X_n)$. In particular, a Bayesian nonparametric estimator of $\tau_{1,q}$, is given by the posterior mean

$$
\begin{aligned}
\hat{\tau}_{1,q}&=\mathbb{E}\{\tau_{1,q}\,|\,X_1,\ldots,X_n\}\\
&=m_1\sum_{i=0}^{q}\frac{(N-n)!}{(N-n-i)!}\frac{(1-\alpha)_{(i)}}{i!}\frac{(\theta+n-1+\alpha)_{(N-n-i)}}{(\theta+n)_{(N-n)}}.
\end{aligned}
$$

Note that $\hat{\tau}_{1,0}$ coincides with $\hat{\tau}_1$ in Theorem 1. We conclude the study of $\tau_{1,q}$ by considering the large n and large N asymptotic behaviour of the posterior (falling) factorial moment $\mathbb{E}\{(\tau_{1,q})_{[r]} \,|\, X_1, \ldots, X_n\}$. In particular, we write the following

$$
\begin{aligned}
&\mathbb{E}\{(\tau_{1,q})_{[z]} \,|\, X_1, \ldots, X_n\} \\
&\quad = z! \binom{m_1}{z} \sum_{i_1=0}^{q} \cdots \sum_{i_z=0}^{q} \\
&\qquad \times \frac{(N-n)!}{(N-n-\sum_{t=1}^{z} i_t)!} \prod_{t=1}^{z} \frac{(1-\alpha)_{(i_t)}}{i_t!} \frac{(\theta+n-z+z\alpha)_{(N-n-\sum_{t=1}^{z} i_t)}}{(\theta+n)_{(N-n)}} \\
&\quad = z! \binom{m_1}{z} \sum_{i_1=0}^{q} \cdots \sum_{i_z=0}^{q} \\
&\qquad \times \frac{\Gamma(N-n+1)/\Gamma(N)}{\Gamma(N-n-\sum_{t=1}^{z} i_t+1)/\Gamma(N)} \prod_{t=1}^{z} \frac{(1-\alpha)_{(i_t)}}{i_t!} \frac{\frac{\Gamma(\theta-z+z\alpha+N-\sum_{t=1}^{z} i_t)/\Gamma(N)}{\Gamma(\Gamma(\theta+n-z+z\alpha)/\Gamma(n)}}{\frac{\Gamma(\theta+N)/\Gamma(N)}{\Gamma(\theta+n)/\Gamma(n)}} \\
&\quad \approx z! \binom{m_1}{z} \sum_{i_1=0}^{q} \cdots \sum_{i_z=0}^{q} \frac{N^{-n+1}}{N^{-n-\sum_{t=1}^{z} i_t+1}} \prod_{t=1}^{z} \frac{(1-\alpha)_{(i_t)}}{i_t!} \frac{\frac{N^{\theta-r+r\alpha-\sum_{t=1}^{z} i_t}}{n^{\theta-z+z\alpha}}}{\frac{N^{\theta}}{n^{\theta}}} \\
&\quad = z! \binom{m_1}{z} \left[\left(\frac{n}{N} \right)^{1-\alpha} \right]^{z} \sum_{i_1=0}^{q} \cdots \sum_{i_z=0}^{q} \prod_{t=1}^{z} \frac{(1-\alpha)_{(i_t)}}{i_t!} \\
&\quad = z! \binom{m_1}{z} \left[\left(\frac{n}{N} \right)^{1-\alpha} \right]^{z} \left[\frac{\Gamma(2+q-\alpha)}{\Gamma(1+q)\Gamma(2-\alpha)} \right]^{z} \\
&\quad = z! \binom{m_1}{z} \left[\left(\frac{n}{N} \right)^{1-\alpha} \frac{\Gamma(2+q-\alpha)}{\Gamma(1+q)\Gamma(2-\alpha)} \right]^{z}. \qquad (30)
\end{aligned}
$$

If

$$
\left(\frac{n}{N} \right)^{1-\alpha} \frac{\Gamma(2+q-\alpha)}{\Gamma(1+q)\Gamma(2-\alpha)} \in (0,1) \qquad (31)
$$

then Equation (30) is the falling factorial moment of order z of a Binomial random variable with parameter $(m_1, (n/N)^{1-\alpha}\Gamma(2+q-\alpha)/\Gamma(1+q)\Gamma(2-\alpha))$, with m_1 being the number of trials and $(n/N)^{1-\alpha}\Gamma(2+q-\alpha)/\Gamma(1+q)\Gamma(2-\alpha)$ being the probability of success in a trial. Note that (31) is always satisfied for $q = 0$.

Acknowledgments

The authors are grateful to an Associate Editor and an anonymous Referee for all their critical comments, corrections, and suggestions which improved remarkably the present paper.

References

[1] Bacallado, S., Battiston, M., Favaro, S. and Trippa, L. (2015). Sufficientness postulates for Gibbs-type priors and hierarchial generalizations. *Statistical Science* **32**, 487–500. MR3730518

[2] Bethlehem, J.G., Keller, W.J. and Pannekoek, J. (1990). Disclosure control of microdata. *Journal of the American Statistical Association* **85** 38–45.

[3] Canale, A., Lijoi, A., Nipoti, B. and Prünster, I. (2017). On the Pitman-Yor process with spike and slab base measure. *Biometrika*, **104** 681–697. MR3694590

[4] Carota, C., Filippone, M., Leombruni, R. and Polettini, S. (2015). Bayesian nonparametric disclosure risk estimation via mixed effects log-linear models. *The Annals of Applied Statistics* **9**, 525–546. MR3341126

[5] Carota, C., Filippone, M. and Polettini, S. (2018). Assessing Bayesian semi-parametric log-linear models: an application to disclosure risk estimation. *International Statistical Review*, to appear.

[6] Camerlenghi, F., Favaro, S., Naulet, Z. and Panero, F. (2020). Optimal disclosure risk assessment. *The Annals of Statistics*, **49**, 723–744. MR4255105

[7] Charalambides, C.A. (2005) *Combinatorial methods in discrete distributions*, Wiley Series in Probability and Statistics. MR2131068

[8] Favaro, S., Lijoi, A. and Prünster, I. (2013). Conditional formulae for Gibbs-type exchangeable random partitions. *The Annals of Applied Probability*, **23**, 1721–1754. MR3114915

[9] Favaro, S., Naulet, Z. (2021). Near-optimal estimation of the unseen under regularly varying tail populations. *arXiv:2104.03251.*

[10] Ferguson, T.S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics* **1**, 209–230. MR0350949

[11] Fienberg, S.E. and Makov, U.E. (1998). Condentiality, uniqueness, and disclosure limitation for categorical data *J. Off. Stat.* **14**, 385–397.

[12] Johnson, N.L., Kemp, A.W. and Kotz, S. (2005) *Univariate discrete distributions*, Wiley Series in Probability and Statistics. MR2163227

[13] Manrique-Vallier, D. and Reiter, J.P. (2012). Estimating identification disclosure risk using mixed membership models. *Journal of the American Statistical Association* **107**, 1385–1394. MR3036402

[14] Manrique-Vallier, D. and Reiter, J.P. (2014). Bayesian estimation of discrete multivariate latent structure models with structural zeros. *Journal of Computational and Graphical Statistics*, **23** 1061–1079. MR3270711

[15] Perman, M., Pitman, J. and Yor, M. (1992). Size-biased sampling of Poisson point processes and excursions. *Probability Theory and Related Fields* **92**, 21–39. MR1156448

[16] Pitman, J. (1995). Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields* **102**, 145–158. MR1337249

[17] Pitman, J. (2006). *Combinatorial Stochastic Processes.* Ecole d'Eté

de Probabilités de Saint-Flour XXXII. Lecture Notes in Mathematics, Springer New York. MR2245368

[18] Pitman, J. and Yor, M. (1999). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability* **25**, 855–900. MR1434129

[19] Reiter, J.P. (2005). Estimating risks of identification disclosure in microdata. *J. Amer. Statist. Assoc.* **100**, 1103–1112. MR2236926

[20] Rinott, Y. and Shlomo, N. (2006). A generalized negative binomial smoothing model for sample disclosure risk estimation. In *Privacy in Statistical Databases. Lecture Notes in Computer Science*, Springer, Berlin. MR2459186

[21] Samuels, S.M. (1998). A Bayesian, species-sampling-inspired approach to the uniques problem in microdata disclosure risk assessment. *Journal of Official Statistics* **14**, 373–383.

[22] Scarpa, B. and Dunson, D. (2009). Bayesian hierarchical functional data analysis via contaminated informative priors. *Biometrics*, **65** 772–780. MR2649850

[23] Skinner, and Elliot, M.J. (2002). A measure of disclosure risk for microdata. *Journal of the Royal Statistical Society, Series B* **64**, 855–867. MR1979391

[24] Skinner, C.J., Marsh, C., Openshaw, S. and Wymer, C. (1994). Disclosure control for census microdata. *Journal of Official Statistics* **10**, 31–51.

[25] Skinner, and Shlomo, N. (2008). Assessing identification risk in survey microdata using log-linear models. *J. Amer. Statist. Assoc.* **103**, 989–1001. MR2462887

[26] Willenborg, L. and de Waal, T. (2001). *Elements of statistical disclosure control.* Springer, New York. MR1866909

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

Title of Paper	**Bayesian nonparametric disclosure risk assessment**
Publication Status	☑ Published □ Accepted for Publication □ Submitted for Publication □ Unpublished and unsubmitted work written in a manuscript style
Publication Details	Stefano Favaro, Francesca Panero, Tommaso Rigon "Bayesian nonparametric disclosure risk assessment", Electronic Journal of Statistics, Electron. J. Statist. 15(2), 5626-5651, (2021)

Student Confirmation

Student Name:	Francesca Panero		
Contribution to the Paper	I am joint second author. I implemented part of the simulation experiments to compare our estimators with others available in the literature. I found the data of the American Community Survey and implemented the comparisons among estimators. I carried out the literature review presented in the introduction.		
Signature		Date	19/04/2022

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title:	Professor François Caron		
Supervisor comments			
Signature		Date	19/04/2022

This completed form should be included in the thesis, at the end of the relevant chapter.

Chapter 5

Optimal disclosure risk assessment

Federico Camerlenghi, Stefano Favaro, Zacharie Naulet, Francesca Panero. "Optimal disclosure risk assessment", The Annals of Statistics, 49(2) 723-744 April 2021.

The Annals of Statistics
2021, Vol. 49, No. 2, 723–744
https://doi.org/10.1214/20-AOS1975

OPTIMAL DISCLOSURE RISK ASSESSMENT

BY FEDERICO CAMERLENGHI[1], STEFANO FAVARO[2], ZACHARIE NAULET[3] AND FRANCESCA PANERO[4]

Dedicated to the memory of Chris Skinner

[1]*Department of Economics, Management and Statistics, University of Milano–Bicocca, federico.camerlenghi@unimib.it*
[2]*Department of Economics and Statistics, Univeristy of Torino, stefano.favaro@unito.it*
[3]*Université Paris-Saclay, Laboratoire de mathématiques d'Orsay, zacharie.naulet@universite-paris-saclay.fr*
[4]*Department of Statistics, University of Oxford, francesca.panero@stats.ox.ac.uk*

Protection against disclosure is a legal and ethical obligation for agencies releasing microdata files for public use. Consider a microdata sample of size n from a finite population of size $\bar{n} = n + \lambda n$, with $\lambda > 0$, such that each sample record contains two disjoint types of information: identifying categorical information and sensitive information. Any decision about releasing data is supported by the estimation of measures of disclosure risk, which are defined as discrete functionals of the number of sample records with a unique combination of values of identifying variables. The most common measure is arguably the number τ_1 of sample unique records that are population uniques. In this paper, we first study nonparametric estimation of τ_1 under the Poisson abundance model for sample records. We introduce a class of linear estimators of τ_1 that are simple, computationally efficient and scalable to massive datasets, and we give uniform theoretical guarantees for them. In particular, we show that they provably estimate τ_1 all of the way up to the sampling fraction $(\lambda + 1)^{-1} \propto (\log n)^{-1}$, with vanishing normalized mean-square error (NMSE) for large n. We then establish a lower bound for the minimax NMSE for the estimation of τ_1, which allows us to show that: (i) $(\lambda + 1)^{-1} \propto (\log n)^{-1}$ is the smallest possible sampling fraction for consistently estimating τ_1; (ii) estimators' NMSE is near optimal, in the sense of matching the minimax lower bound, for large n. This is the main result of our paper, and it provides a rigorous answer to an open question about the feasibility of nonparametric estimation of τ_1 under the Poisson abundance model and for a sampling fraction $(\lambda + 1)^{-1} < 1/2$.

1. Introduction. Protection against disclosure is a legal and ethical obligation for agencies releasing microdata files for public use. Any decision about release requires a careful assessment of the risk of disclosure, which is supported by the estimation of measures of disclosure risk ([30]). Let consider a microdata sample $\boldsymbol{X}(n) = (X_1, \ldots, X_n)$ from a finite population of size $\bar{n} > n$ and, without loss of generality, assume that each X_i is a record containing two disjoint types of information for the ith individual: identifying information and sensitive information. Identifying information consists of a set of categorical variables which might be matchable to known units of the population. A risk of disclosure results from the possibility that an intruder might succeed in identifying a microdata unit through such a matching, and hence be able to disclose sensitive information on this unit. To quantify the risk of disclosure, sample records $\boldsymbol{X}(n)$ are typically cross-classified according to identifying variables. That is, $\boldsymbol{X}(n)$ is partitioned in $K_n \leq n$ cells, with $Y_j(\boldsymbol{X}, n)$ being the number of

Received February 2019; revised April 2020.
MSC2020 subject classifications. Primary 62G05; secondary 62C20.
Key words and phrases. Disclosure risk assessment, microdata sample, nonparametric inference, optimal minimax procedure, Poisson abundance model, polynomial approximation.

X_i's belonging to cell j, for $j = 1, \ldots, K_n$, such that $\sum_{1 \le j \le K_n} Y_j(\boldsymbol{X}, n) = n$; we refer to the number of occurrences $Y_j(\boldsymbol{X}, n)$ as the sample frequency of cell j. Then a risk of disclosure arises from cells in which both sample frequencies and population frequencies are small. Of special interest are cells with frequency 1 (singletons or uniques) since, assuming no errors in the matching process or data sources, for these cells the match is guaranteed to be correct. This has motivated inferences on measures of disclosure risk that are suitable functionals of the number of uniques, the most common being the number τ_1 of sample uniques which are also population uniques. We refer to [24] for a comprehensive account on measures of disclosure risk.

In this paper, we first study nonparametric estimation of the discrete functional τ_1 under the Poisson abundance model for sample records. The Poisson abundance model is arguably the most natural, and weak, assumption to infer τ_1 ([2] and [25]). If $\bar{n} = n + \lambda n$, with $\lambda > 0$, the model assumes that: (i) the population records $(X_1, \ldots, X_{n+\lambda n})$ can be ideally extended to a sequence $\boldsymbol{X} = (X_i)_{i \ge 1}$, of which $\boldsymbol{X}(n)$ is an observable subsample; (ii) the X_i's are independent and identically distributed as an unknown distribution $(p_j)_{j \ge 1}$, where p_j is the probability of the jth cell in which $\boldsymbol{X}$ may be cross-classified; (iii) the sample size is a Poisson random variable N with mean n, in symbols $N \sim \text{Poiss}(n)$. Then sample records $\boldsymbol{X}(N) = (X_1, \ldots, X_N)$ result in K_N cells with $Y_j(\boldsymbol{X}, N)$ being the sample frequency of cell j, for $j = 1, \ldots, K_N$, such $Y_j(\boldsymbol{X}, N) \sim \text{Poiss}(np_j)$, $Y_{j_1}(\boldsymbol{X}, N)$ is independent of $Y_{j_2}(\boldsymbol{X}, N)$ for any $j_1 \ne j_2$, and $\sum_{1 \le j \le K_N} Y_j(\boldsymbol{X}, N) = N$. [26] first raised the problem of nonparametric estimation of τ_1 under the Poisson abundance model, leaving that as an open problem. In particular, they discussed about the feasibility of nonparametric estimation of τ_1, arguing that it is an intrinsically difficult problem. The problem shares the well-known difficulties of the classical problem of estimating the number of unseen species ([10] and [8]). Indeed nonparametric estimators of τ_1 may be "unreasonable" since they are subject to serious upward bias and high variance for small sampling fractions of the population, that is, $(\lambda + 1)^{-1} < 1/2$ or, in other words, for n smaller than a half of the population $\bar{n}$.

Under the Poisson abundance model for sample records $\boldsymbol{X}(n)$ from the population $(X_1, \ldots, X_{n+\lambda n})$, we introduce a class of nonparametric linear estimators of τ_1 that are simple, computationally efficient and scalable to massive datasets. We show that our estimators admit an interpretation as (smoothed) nonparametric empirical Bayes estimators in the sense of [22], and we prove theoretical guarantees for them that hold uniformly for any distribution $(p_j)_{j \ge 1}$. In particular, we show that our estimators provably estimate τ_1 all of the way up to the sampling fraction $(\lambda + 1)^{-1} \propto (\log n)^{-1}$ of the population, with vanishing normalized mean-square error (NMSE) as n becomes large. Then, by relying on recent techniques developed in [32] in the context of nonparametric estimation of the support size of discrete distributions, we establish a lower bound for the minimax NMSE for the estimation of τ_1. This result allows us to show that $(\lambda + 1)^{-1} \propto (\log n)^{-1}$ is the smallest possible sampling fraction of the population for consistently estimating τ_1, and that the estimators' NMSE is near optimal, in the sense of matching the minimax lower bound, for a large sample size n. This is the main result of the present paper, and it provides a rigorous answer to the question raised by [26] about the feasibility of nonparametric estimation of τ_1 under the Poisson abundance model and for a sampling fraction $(\lambda + 1)^{-1} < 1/2$. Indeed our result shows that nonparametric estimation of τ_1 has uniformly provable guarantees, in terms of vanishing NMSE for large n, if and only if $(\lambda + 1)^{-1} \propto (\log n)^{-1}$.

Starting from the seminal work of [2], in the last three decades a full range of parametric and semiparametric approaches, both frequentist and Bayesian, has been proposed for making inference on τ_1; see, for example, [4, 13, 14, 20, 21, 23–25] and [5]. A common thread of these works has been the enrichment of the classical Poisson abundance model with stronger modeling assumptions: while early approaches were focused on parametric

Bayesian modeling of the random partition induced by the cross classification of sample records, recent approaches focused on semiparametric modeling of the associations among identifying variables, typically by means of complex Bayesian hierarchical latent class models. All approaches in the literature are shown to empirically estimate τ_1, even for relatively small sampling fractions, but without any provable guarantees. The approach we propose in the present paper may be viewed as the natural nonparametric counterpart of the parametric empirical Bayes approach, in the sense of [7], introduced in [2] and further developed in [24] and [21]. Besides being the first nonparametric approach to the estimation of τ_1 under the Poisson abundance model, our approach stands out for being the first to give theoretical guarantees on the performance of the proposed class of estimators.

The paper is structured as follows. In Section 2, we introduce a class of nonparametric estimators for τ_1, and we show that they provably estimate τ_1 all of the way up to the sampling fraction $(\lambda+1)^{-1} \propto (\log n)^{-1}$, with vanishing NMSE for large sample size n. In Section 3, we show that $(\lambda+1)^{-1} \propto (\log n)^{-1}$ is the smallest possible sampling fraction of the population which guarantees a vanishing NMSE, and that estimators' NMSE is near optimal for large n. Section 4 contains a discussion of our results, their interplay with other discrete functional estimation problems, and remaining open challenges. Proofs are deferred to the Appendix, whereas technical results and numerical illustrations are available as Supplementary Material [3].

2. A nonparametric estimator of τ_1. We consider an infinite sequence of observations $\boldsymbol{X}$, and we assume that $\boldsymbol{X}(N) = (X_1, \ldots, X_N)$ is the microdata sample of random size N under the Poisson abundance model. We suppose that $\boldsymbol{X}(N)$ is a subsample of $(X_1, \ldots, X_{M+N})$, where $M \sim \text{Poiss}(\lambda n)$, with $\lambda > 0$ and independent of N. In the present framework, $(X_{N+1}, \ldots, X_{N+M})$ may be seen as the unobservable population. When sample records are cross-classified according to identifying variables, the sample $(X_1, \ldots, X_N)$ results partitioned in $K_N \leq N$ cells with corresponding sample frequencies $(Y_1(\boldsymbol{X}, N), \ldots, Y_{K_N}(\boldsymbol{X}, N))$ such that $\sum_{1 \leq j \leq K_N} Y_j(\boldsymbol{X}, N) = N$. Hereafter, we denote by $Z_i(\boldsymbol{X}, N)$ the number of cells with frequency i, and by $Z_{\tilde{i}}(\boldsymbol{X}, N)$ the number of cells with frequency greater or equal than i, for any index $i \geq 1$. We are interested in estimating the number τ_1 of sample uniques which are also population uniques, namely the following discrete functional:

$$\tau_1(\boldsymbol{X}, N, M) = \sum_{j \geq 1} \mathbb{1}_{\{Y_j(\boldsymbol{X}, N)=1\}} \mathbb{1}_{\{Y_j(\boldsymbol{X}, N+M)=1\}},$$

where $\mathbb{1}$ denotes the indicator function. We recall that the frequency counts, defined as $Y_j(\boldsymbol{X}, N) = \sum_{1 \leq i \leq N} \mathbb{1}_{\{X_i = j\}}$, are distributed according to a Poisson distribution with parameter np_j, where p_j is the unknown probability associated to the jth cell, that is, $p_j \in [0, 1]$ for $j \geq 1$ such that $\sum_{j \geq 1} p_j = 1$. We will denote by $\boldsymbol{Y}(\boldsymbol{X}, N) := (Y_1(\boldsymbol{X}, N), \ldots)$ the whole sequence of the cell's frequency counts. We remark that, under the Poisson abundance model, the $Y_j(\boldsymbol{X}, N)$'s are independent random variables variables and, in addition, $Y_j(\boldsymbol{X}, N+M) - Y_j(\boldsymbol{X}, N)$ is independent of $Y_j(\boldsymbol{X}, N)$, for any $j \geq 1$: these properties follow from standard statistical arguments. When the sample size n is fixed, the independence property of the $Y_j(\boldsymbol{X}, n)$'s falls down and approximation arguments are required to handle such a situation.

To fix the notation, in the sequel we will write $f \lesssim g$, for two generic functions f and g, if and only if (iff) there exists a universal constant $C > 0$ such that $f(x) \leq Cg(x)$; we will further write $f \asymp g$ whenever both $f \lesssim g$ and $g \lesssim f$ are satisfied. Let us denote by $\mathscr{P}$ the set of all possible distributions over the set of natural numbers $\mathbb{N}$, that is, $\mathscr{P} := \{P = \sum_{j \geq 1} p_j \delta_j : p_j \in [0, 1], \text{ with } \sum_{j \geq 1} p_j = 1\}$, where δ_j denotes the Dirac measure centered at $j \in \mathbb{N}$. An estimator of $\tau_1(\boldsymbol{X}, N, M)$ is understood to be a measurable function $\hat{\rho}_1(\boldsymbol{X}(N), N)$

depending on the available sample $\boldsymbol{X}(N)$ and the actual size of the observed sample N. We will evaluate the performance of a generic estimator $\hat{\rho}_1(\boldsymbol{X}(N), N)$ of $\tau_1(\boldsymbol{X}, N, M)$, by its worst-case NMSE, defined as

$$\mathscr{E}_{\lambda,n}\big(\hat{\rho}_1\big(\boldsymbol{X}(N), N\big)\big) := \sup_{P\in\mathscr{P}} \frac{\mathbb{E}[(\hat{\rho}_1(\boldsymbol{X}(N), N) - \tau_1(\boldsymbol{X}, N, M))^2]}{n^2}, \tag{1}$$

where $\mathbb{E}[(\hat{\rho}_1(\boldsymbol{X}(N), N) - \tau_1(\boldsymbol{X}, N, M))^2]$ is the mean squared error (MSE) of $\hat{\rho}_1$ under the model (P, n, λ), also denoted by $\mathrm{MSE}[\hat{\rho}_1(\boldsymbol{X}(N), N)]$. Since $\mathrm{MSE}[\hat{\rho}_1(\boldsymbol{X}(N), N)]$ does not vanish as $n \to +\infty$, it is common to evaluate the performance of an estimator for $\tau_1(\boldsymbol{X}, N, M)$ in terms of the NMSE (see, e.g., [19]). The NMSE is indeed the MSE of $\hat{\rho}_1(\boldsymbol{X}(N), N)$ normalized by the maximum value of $\tau_1(\boldsymbol{X}, N, M)$ (which is exactly n, given $N = n$), and hence the performance of $\hat{\rho}_1(\boldsymbol{X}(N), N)$ is evaluated in terms of the rate of convergence to 0 of the NMSE as $n \to +\infty$.

A nonparametric estimator for $\tau_1(\boldsymbol{X}, N, M)$ may be simply deduced by comparing expectations. Indeed, under the Poisson abundance model, it easy to see that

$$\mathbb{E}\big[\tau_1(\boldsymbol{X}, N, M)\big] = \sum_{i\geq 0}(-1)^i\lambda^i(i+1)\mathbb{E}\big[Z_{i+1}(\boldsymbol{X}, N)\big]. \tag{2}$$

See Appendix A.1 for details on the derivation of identity (2). In particular, according to identity (2) we can define the following estimator of $\tau_1(\boldsymbol{X}, N, M)$:

$$\hat{\tau}_1\big(\boldsymbol{X}(N), N\big) = \sum_{i\geq 0}(-1)^i(i+1)\lambda^i Z_{i+1}(\boldsymbol{X}, N). \tag{3}$$

By construction $\hat{\tau}_1(\boldsymbol{X}(N), N)$ is an unbiased estimator of $\mathbb{E}[\tau_1(\boldsymbol{X}, N, M)]$, that is, $\mathbb{E}[\hat{\tau}_1(\boldsymbol{X}(N), N)] = \mathbb{E}[\tau_1(\boldsymbol{X}, N, M)] = \sum_{j\geq 1} np_j e^{-(\lambda+1)np_j}$. The estimator $\hat{\tau}_1(\boldsymbol{X}(N), N)$ admits a natural interpretation as a nonparametric empirical Bayes estimator in the sense of [22]. More precisely, $\hat{\tau}_1(\boldsymbol{X}(N), N)$ is the posterior expectation of $\mathbb{E}[\tau_1(\boldsymbol{X}, N, M)]$ with respect to an unknown prior distribution on the p_i's that is estimated from the $Y_j(\mathbf{X}, N)$. See Appendix A.2 for details. This observation makes the estimator (3) the natural nonparametric counterpart of the parametric empirical Bayes estimator, in the sense of [7], introduced in [2].

THEOREM 1. *For any positive reals x and y, let $\lfloor x\rfloor$ denote the integer part of x and let $x\vee y$ denote the maximum between x and y. If $\lambda < 1$, for any $P\in\mathscr{P}$ and for any $n > 0$,*

$$\begin{aligned}&\mathrm{Var}\big[\tau_1(\boldsymbol{X}, N, M) - \hat{\tau}_1\big(\boldsymbol{X}(N), N\big)\big]\\ &\quad\leq \Psi^2(\lambda)\mathbb{E}\big[Z_{\bar{1}}(\boldsymbol{X}, N)\big] - \frac{\mathbb{E}[Z_1(\boldsymbol{X}, N+M)]}{\lambda+1},\end{aligned} \tag{4}$$

where in (4) we defined $\Psi(\lambda) = (j^*+1)\lambda^{j^*}$ such that $j^* = \lfloor(2\lambda-1)/(1-\lambda)\rfloor\vee 0$.

The proof of Theorem 1 is deferred to Appendix A.3. According to Theorem 1, for $\lambda < 1$ one has that $\mathrm{Var}[\tau_1(\boldsymbol{X}, N, M) - \hat{\tau}_1(\boldsymbol{X}(N), N)] \lesssim n$ upon noticing that $\mathbb{E}[Z_{\bar{1}}(\boldsymbol{X}, N)] \leq \mathbb{E}[N] = n$. That is, in expectation, $\hat{\tau}_1(\boldsymbol{X}(N), N)$ approximate $\tau_1(\boldsymbol{X}, N, M)$ to within n. We formalize these observations in the next corollary.

COROLLARY 1. *If $\lambda < 1$ is fixed, then $\mathscr{E}_{\lambda,n}(\hat{\tau}_1(\boldsymbol{X}(N), N)) \leq W(\lambda)/n$, for any $n \geq 1$ and for some constant $W(\lambda)$ depending only on λ.*

Corollary 1 legitimates $\hat{\tau}_1(\boldsymbol{X}(N), N)$ as an estimator of $\tau_1(\boldsymbol{X}, N, M)$ under the assumption $\lambda < 1$. Unfortunately, this assumption is unrealistic in the context of disclosure risk assessment, where the size λn of the unobserved population is typically much bigger than the size n

of the observed sample. The variance bound in Theorem 1 reveals that the assumption $\lambda < 1$ is necessary to obtain a finite estimate of the variance. This variance issue of $\hat{\tau}_1(\boldsymbol{X}(N), N)$ is determined by the geometrically increasing magnitude of the coefficients $(i+1)(-\lambda)^i$. Indeed, as $\lambda \geq 1$, the estimator $\hat{\tau}_1(\boldsymbol{X}(N), N)$ grows superlinearly as $(i+1)(-\lambda)^i$ for the largest i such that $Z_{i+1}(\boldsymbol{X}, N) > 0$, thus eventually far exceeding $\tau_1(\boldsymbol{X}, N, M)$ that grows at most linearly. Then $\hat{\tau}_1(\boldsymbol{X}(N), N)$ is useless for $\lambda \geq 1$, thus requiring an adjustment via suitable smoothing techniques. To fix this issue, we follow ideas developed by [8, 10] and [19] in the context of the nonparametric estimation of the number of unseen species. We propose a smoothed version of $\hat{\tau}_1(\boldsymbol{X}(N), N)$ by truncating the series (3) at an independent random location L, and then averaging over the distribution of L, that is,

$$\begin{aligned}\hat{\tau}_1^L\big(\boldsymbol{X}(N), N\big) &= \mathbb{E}_L\left[\sum_{i=1}^{L}(-1)^i(i+1)\lambda^i Z_{i+1}(\boldsymbol{X}, N)\right] \\ &= \sum_{i\geq 0}(-1)^i(i+1)\lambda^i \mathbb{P}(L \geq i) Z_{i+1}(\boldsymbol{X}, N).\end{aligned} \tag{5}$$

For any $\lambda \geq 1$, as the the index i in (5) increases, the tail probability $\mathbb{P}[L \geq j]$ compensate for the exponential growth of $(i+1)(-\lambda)^i$, thereby stabilizing the variance. In the next theorem, we show that for $\lambda \geq 1$ the estimator $\hat{\tau}_1^L(\boldsymbol{X}(N), N)$ is biased for $\mathbb{E}[\tau_1(\boldsymbol{X}, N, M)]$, and we provide a bound for the MSE of $\hat{\tau}_1(\boldsymbol{X}(N), N)$.

THEOREM 2. *Let $\hat{\tau}_1^L(\boldsymbol{X}(N), N)$ be the estimator of $\tau_1(\mathbf{X}, N, M)$ defined in (5). If $\lambda \geq 1$, then*

$$\begin{aligned}&\mathbb{E}\big[\hat{\tau}_1^L\big(\boldsymbol{X}(N), N\big)\big] \\ &\quad = \mathbb{E}\big[\tau_1(\boldsymbol{X}, N, M)\big] + \sum_{j\geq 1} e^{-p_j n(\lambda+1)} p_j n \int_0^{\lambda n p_j} e^s \mathbb{E}_L\left[\frac{(-s)^L}{L!}\right] \mathrm{d}s\end{aligned} \tag{6}$$

and

$$\begin{aligned}&\mathrm{MSE}\big[\hat{\tau}_1^L\big(\boldsymbol{X}(N), N\big)\big] \\ &\quad \leq \big(\mathbb{E}_L\big[(L+1)\lambda^L\big]\big)^2 \mathbb{E}\big[Z_{\tilde{1}}(\boldsymbol{X}, N)\big] - \frac{\mathbb{E}[Z_1(\boldsymbol{X}, N+M)]}{\lambda+1} \\ &\qquad + \left(\sum_{j\geq 1} e^{-p_j n(\lambda+1)} p_j n \int_0^{\lambda n p_j} e^s \mathbb{E}_L\left[\frac{(-s)^L}{L!}\right] \mathrm{d}s\right)^2.\end{aligned} \tag{7}$$

The proof of Theorem 2 is in Appendix A.4. Choosing different smoothing distributions for L yields different estimators for $\tau_1(\boldsymbol{X}, N, M)$. Following [19], we consider two distributions for L: (i) a Poisson distribution with parameter $\beta > 0$; (ii) a Binomial distribution with parameter $(x_0, 2/(\lambda+2))$. To choose the parameter β of the Poisson distribution and the parameter x_0 of the Binomial distribution, one should look for $\tilde{\beta}$ and $\tilde{x}_0$ which minimizes the MSE bound (7). Once the values of $\tilde{\beta}$ and $\tilde{x}_0$ are determined explicitly, we are able to obtain a "limit of predictability" for $\hat{\tau}_1^L(\boldsymbol{X}(N), N)$. That is, for some $\delta > 0$ we are able to specify the maximum value of the sampling fraction λ for which $\mathscr{E}_{\lambda,n}(\hat{\tau}_1^L(\boldsymbol{X}(N), N)) < \delta$. This gives a provable (performance) guarantee for the estimation of $\tau_1(\boldsymbol{X}, N, M)$ in terms of λ.

PROPOSITION 1. *Let L be a Poisson random variable with parameter β. Then*

$$\mathrm{MSE}\big[\hat{\tau}_1^L\big(\boldsymbol{X}(N), N\big)\big] \leq e^{-2\beta} n^2 + n e^{2\beta(2\lambda-1)}. \tag{8}$$

The right-hand side of (8) *in minimized by setting* $\tilde{\beta} = \log(n/(2\lambda - 1))/(4\lambda)$, *for any* $\lambda \geq 1$. *Moreover, if* L *is a Poisson random variable with parameter* $\tilde{\beta}$ *then*

$$\mathscr{E}_{n,\lambda}\big(\hat{\tau}_1^L\big(\boldsymbol{X}(N), N\big)\big) \leq \frac{A(\lambda)}{n^{1/(2\lambda)}}, \tag{9}$$

and for any $\delta \in (0, 1)$,

$$\lim_{n\to+\infty} \frac{\max\{\lambda : \mathscr{E}_{n,\lambda}(\hat{\tau}_1^L(\boldsymbol{X}(N), N)) \leq \delta\}}{\log(n)} \geq \frac{1}{2\log(A/\delta)}, \tag{10}$$

where $A(\lambda)$, *is continuous in* $[1, +\infty)$ *with* $\lim_{\lambda\to+\infty} A(\lambda) = 1$ *and* $A = \max_{\lambda\geq 1} A(\lambda) < +\infty$.

See Appendix A.5 for the proof of Proposition 1. A result similar to Proposition 1 holds true when the random variable L is assumed to be distributed according to a Binomial distribution. This result is stated in the next proposition, and its proof is omitted since it is along lines similar to the proof of Proposition 1.

PROPOSITION 2. *For any positive reals* x *and* y, *let* $\lfloor x \rfloor$ *denote the integer part of* x. *Let* L *be a Binomial random variable with parameter* $(x_0, 2/(\lambda + 2))$. *Then*

$$\mathrm{MSE}\big[\hat{\tau}_1^L(\boldsymbol{X}(N), N)\big] \leq n\left(\frac{\lambda}{\lambda+2}\right)^{2x_0}\left[3^{10x_0/3} + n\left(\frac{\lambda}{2(\lambda+1)}\right)^2\right] \tag{11}$$

and the choice $\tilde{x}_0 = \lfloor (3/10)\log_3(n\lambda^2/((\lambda+1)(\lambda^2(3^{10/3}-1) - 4\lambda - 4)))\rfloor$ *minimizes the right-hand side of* (11), *for any* $\lambda \geq 1$. *Moreover, if* L *is a Binomial random variable with parameter* $(\tilde{x}_0, 2/(\lambda+2))$ *then*

$$\mathscr{E}_{n,\lambda}\big(\hat{\tau}_1^L\big(\boldsymbol{X}(N), N\big)\big) \leq \frac{C(\lambda)}{n^{3\log_3(1+2/\lambda)/5}}, \tag{12}$$

and for any $\delta \in (0, 1)$,

$$\lim_{n\to+\infty} \frac{\max\{\lambda : \mathscr{E}_{n,\lambda}(\hat{\tau}_1^L) \leq \delta\}}{\log(n)} \geq \frac{6}{5\log(3)\log(C/\delta)}, \tag{13}$$

where $C(\lambda)$ *is continuous in* $[1, +\infty)$ *with* $\lim_{\lambda\to+\infty} C(\lambda) = 1$ *and* $C = \max_{\lambda\geq 1} C(\lambda)$.

3. Optimality of the proposed estimators. In Section 2, we have introduced two different estimators of $\tau_1(\boldsymbol{X}, N, M)$, and we have provided guarantees of their performance, as $n \to +\infty$, in terms of the NMSE. We have already remarked that the case $\lambda \geq 1$ is the most interesting one for estimating the disclosure risk $\tau_1(\boldsymbol{X}, N, M)$. Indeed in the context of disclosure risk assessment the fraction of the unobserved sample λ is usually much larger than 1. Throughout the section, we assume that $\lambda \geq 1$ and we prove that the proposed estimator $\hat{\tau}_1^L(\boldsymbol{X}(N), N)$ is essentially optimal. More precisely, we determine a lower bound for the best worst-case NMSE, defined as

$$\mathscr{E}(\lambda, n) := \inf_{\hat{\rho}_1} \mathscr{E}_{\lambda,n}\big(\hat{\rho}_1\big(\boldsymbol{X}(N), N\big)\big), \tag{14}$$

where the infimum in the previous definition runs over all possible estimators $\hat{\rho}_1$ of $\tau_1(\boldsymbol{X}, N, M)$. We will then see that the determined lower bound essentially matches with the upper bound (9). In the sequel, we refer to $\mathscr{E}(\lambda, n)$ as the (normalized) minimax risk. The theorem provides us with a lower bound for $\mathscr{E}(\lambda, n)$.

THEOREM 3. *Assume that* $\liminf_{n\to+\infty}(1+\lambda) > e^2$. *Then there exists a universal constant* $K > 0$ *such that, for any* n *sufficiently large, we have that*

$$\mathscr{E}(\lambda, n) \geq K \cdot \begin{cases} 1 & \text{if } \lambda + 1 > \log(n), \\ \dfrac{1+\lambda}{\log(n)} \left(\dfrac{\sqrt{\log(n)}}{n(1+\lambda)} \right)^{e^2/(1+\lambda)} & \text{if } \lambda + 1 \leq \log(n). \end{cases} \tag{15}$$

According to Theorem 3, it is clear that the lower bound on the (normalized) minimax risk goes to zero if $\lambda + 1 = o(\log(n))$ and the rate is provided by the following corollary.

COROLLARY 2. *Assume that* $1 + \lambda > e^2$. *Then there exist universal constants* $c > 0$ *and* $c' > 0$ *such that, for any* n *sufficiently large, we have that*

$$\mathscr{E}(\lambda, n) \geq c \frac{1}{n^{c'/\lambda}}. \tag{16}$$

Corollary 2 is a consequence of Theorem 3, indeed, when $\lambda + 1 > \log(n)$ the two lower bounds in (15) and (16) are constants, whereas if $\lambda + 1 \leq \log(n)$ it is easy to observe that the leading term in (15), as $n \to +\infty$, is of order $1/n^{c'/\lambda}$ as in (16) for some $c' > 0$. One may easily see that every constant $c' > e^2$ works in (16). Corollary 2 provides us with a lower bound for the NMSE of any estimator of the disclosure risk $\tau_1(\boldsymbol{X}, N, M)$. The lower bound (16) has an important implication: without imposing any parametric assumption on the model, one can estimate $\tau_1(\boldsymbol{X}, N, M)$ with vanishing NMSE all the way up to $\lambda \propto \log n$. It is then impossible to determine an estimator having provable guarantees, in terms of vanishing NMSE, when $\lambda = \lambda(n)$ goes to $+\infty$ much faster than $\log(n)$, as a function of n. By the "limit of predictability" (10) determined for the estimator $\hat{\tau}_1^L(\boldsymbol{X}(N), N)$, we conclude that the proposed estimator is optimal, because its "limit of predictability" matches (asymptotically) with its maximum possible value $\lambda \propto \log(n)$.

3.1. *Guideline for the proof of Theorem 3.* We present the main ingredients for the proof of Theorem 3. Hereafter, we will write $\mathbb{E}_P^{n,\lambda}$ (resp., $\mathbb{P}_P^{n,\lambda}$) in order to make explicit the dependence of the expected value (resp., the probability measure) w.r.t. P, the parameter n of the Poisson random variable N and λ. The proof of Theorem 3 relies on the method of the two fuzzy hypotheses ([28]), which allows to reduce the proof of Theorem 3 to the problem of finding the best polynomial approximation to some functions. A similar approach has been recently considered by [31, 32] in the context of nonparametric estimation of the support size of discrete distributions. Some steps of the proof of Theorem 3 are similar to that of [31] and, therefore, they are omitted here, in favor of highlighting only the key differences. For the sake of completeness, the whole proof is offered in the Supplementary Material [3].

Lemma 1 and Lemma 2 below are used in the proof of Theorem 3, and they constitutes the essential difference between the proof of Theorem 3 and the proof of the minimax lower bound in the work of [31]. Lemma 1 and Lemma 2 are proved in Appendix B.1 and Appendix B.2, respectively.

LEMMA 1. *The following identity holds true*:

$$\mathscr{E}(\lambda, n) = \inf_{\hat{\rho}} \sup_{P \in \mathscr{P}} n^{-2} \mathbb{E}_P^{n,\lambda} \left[\left(\tau_1(\boldsymbol{X}, N, M) - \hat{\rho}(\boldsymbol{Y}(\boldsymbol{X}, N)) \right)^2 \right],$$

where the infinimum in the previous equation is understood to be taken with respect to all measurable maps $\hat{\rho} : \mathbb{N}^{\mathbb{N}} \to \mathbb{R}$.

The definition of the minimax risk in (14) allows for estimators depending on the whole sample $\boldsymbol{X}(N)$, while $\tau_1(\boldsymbol{X}, N, M)$ depends only on the frequencies $\boldsymbol{Y}(\boldsymbol{X}, N+M)$ and $\boldsymbol{Y}(\boldsymbol{X}, N)$. Thus, in view of Lemma 1, there should be no gain of information in using estimators depending on $\boldsymbol{X}(N)$ over estimators depending only on the frequencies $\boldsymbol{Y}(\boldsymbol{X}, N)$. Investigation of the proof of Lemma 1 shows that for all estimators $\hat{\tau}_1$, the estimator $\hat{\rho}$ obtained by symmetrizing $\hat{\tau}_1$ and taking the expectation conditional on $\boldsymbol{Y}(\boldsymbol{X}, N)$ has always risk smaller or equal than $\hat{\tau}_1$. This may be viewed as a form of *Rao–Blackwellisation* of $\hat{\tau}_1$, where $\boldsymbol{Y}(\boldsymbol{X}, N)$ acts as a sufficient statistics for τ_1, in the sense that $\hat{\rho}$ never depends on the distribution of $\boldsymbol{X}$. Besides being of self-interest for the reasons previously invoked, Lemma 1 crucially makes the proof of Theorem 3 easier by remarking that $(\boldsymbol{X}, k) \mapsto \boldsymbol{Y}(\boldsymbol{X}, k)$ is nicely distributed under the Poisson model. The Lemma 1 constitutes the starting point of the proof of Theorem 3. The rest of the proof consists on applying the reduction scheme of [31, 32] to the expression in Lemma 1. The major difference with the aforementioned paper is that we have to find the best, uniform on some interval, polynomial approximation of the map $x \mapsto \exp(-2Bx)$ for arbitrary $B > 0$ instead of the map $x \mapsto \log(x)$ considered in [31].

To be more precise, for $a, b \in \mathbb{R}$, we let $\mathsf{C}[a, b]$ denote the space of continuous functions on $[a, b]$, and for any $L \in \mathbb{Z}_+$ we let $\mathsf{P}_L[a, b] \subset \mathsf{C}[a, b]$ denote the space of polynomials of degree no more than L on $[a, b]$. For any $f \in \mathsf{C}[a, b]$, the best polynomial (of degree at most L) approximation to f is defined as

$$E_L(f, [a, b]) := \inf\{\sup\{|f(x) - q(x)| : x \in [a, b]\} : q \in \mathsf{P}_L[a, b]\}. \tag{17}$$

Then our main result on the best, uniform on some interval, polynomial approximation of the of the map $x \mapsto \exp(-2Bx)$, is stated in the following lemma, proved in Appendix B.2. The rate of approximation is given in term of the function $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ such that

$$\varphi(x) := 1 - \sqrt{1 + x^2} + x \operatorname{arcsinh}(x). \tag{18}$$

LEMMA 2. *Let $\xi > 1$ and $g_\xi : [\xi^{-1}, 1] \to \mathbb{R}_+$ be such that $g_\xi(x) := \exp\{-2B_\xi x\}$ with $B_\xi = (\xi/2)(1 + O(\xi^{-1}))$ as $\xi \to \infty$. Then, for every $\zeta > 0$, there exist constants $K, \xi_0 > 0$ such that for all $\xi > \xi_0$ and all $0 < L \le \zeta\xi$,*

$$E_L(g_\xi, [\xi^{-1}, 1]) \ge K \cdot \begin{cases} 1 & \text{if } 0 < L \le \sqrt{\xi/2}, \\ \dfrac{\sqrt{\xi}}{L} \exp\left\{-\dfrac{\xi}{2}\varphi\left(\dfrac{2L}{\xi}\right)\right\} & \text{if } \sqrt{\xi/2} < L < \zeta\xi. \end{cases}$$

It is worth discussing how the previous result can be of interest beyond its use in this paper. Approximation theory usually focuses on the regime $L/\xi \to \infty$, where the error of approximation is known to be superexponential in L. This regime is omitted here since it is a classical result and we only need the regime $L/\xi \to \gamma$ for some constant $\gamma \ge 0$ in the proof of Theorem 3. Approximation in the latter regime is much more difficult, as emphasized by Lemma 2, and was not studied before to the best of our knowledge. The proof of Lemma 2 uses the well-known duality between best polynomial approximation and best trigonometric polynomial approximation. Using the orthogonality of trigonometric polynomials, we are able to reduce the problem into finding a good lower bound on $\max_{K \in \mathbb{N}} K e^{-C} I_{L+4K}(C)$, where I_k are the modified Bessel function of the first kind (see [18], p. 248), and $C \approx \xi/2$. Then the most delicate and final step consist of establishing the double asymptotic of $I_k(C)$ as $k \to \infty$ and $C \to \infty$, with the constraint that $\sqrt{C} \le k \lesssim C$. Finally, we note that the lower bound in Lemma 2 is essentially sharp, that is, up to determining the value of the constant K. The matching upper-bound is derived in the Supplementary Material [3] by analyzing the rate of convergence of Chebychev polynomials approximation of increasing orders.

4. Discussion. [26] first raised the problem of nonparametric estimation of τ_1 under the Poisson abundance model for sample records, and they left that as an open problem. In this paper, we first considered the problem of [26], and we presented a rigorous solution to it. In particular, we introduced a class of nonparametric estimators of τ_1, and we gave uniform theoretical guarantees for them. First, we showed that our estimators provably estimate τ_1 all of the way up to the sampling fraction $(\lambda + 1)^{-1} \propto (\log n)^{-1}$, with vanishing NMSE as n becomes large. Second, and most importantly, we proved that: (i) $(\lambda + 1)^{-1} \propto (\log n)^{-1}$ is the smallest possible sampling fraction of the population for consistently estimating τ_1; (ii) estimators' NMSE is near optimal, in the sense of matching the minimax lower bound, for large n. Besides being the first study on nonparametric inference for τ_1 under the Poisson abundance model, our work is the first to provide theoretical guarantees on the estimation of τ_1. Indeed, despite the large number of contributions to the estimation of τ_1, all of them proposed parametric and semiparametric approaches that empirically estimate τ_1, but without provable guarantees. To be best of our knowledge, none of the contributions considers a rigorous study on the interplay between the estimation of τ_1 and λ.

The problem of estimating τ_1 belongs to a broad class of discrete functional estimation problems, commonly known as species sampling problems. Consider a population of individuals $(X_i)_{i\geq 1}$ belonging to different "species" $(S_j)_{j\geq 1}$ with unknown proportions $(p_j)_{j\geq 1}$. Given an initial observable samples of size n from the population, species sampling problems refer to the estimation of features of the population or features of λn additional unobservable samples. Recent noteworthy works on species sampling problems are concerned with the estimation of the following discrete functionals: support size (e.g., [29] and [32]); entropy (e.g., [11] and [31]); missing mass (e.g., [15, 17] and [1]); number of unseen species (e.g., [8] and [19]). Interest in these quantities first appeared in ecology, and it has grown in the recent years driven by challenging applications in biosciences, physical sciences, machine learning, engineering, theoretical computer science, information theory, etc. Our study on τ_1 contributes to these recent literature, by studying a new discrete functional of interest in the context of disclosure risk assessment.

While τ_1 is known to be the most common measure of disclosure risk ([2] and [24]), one might consider alternative measures by broadening the definition of "uniqueness." For instance, [9] considered a measure of disclosure risk defined in terms of the number of cells with frequency less or equal than 2. In general, one may consider

$$\tau_{r_N,r_M}(\boldsymbol{X}, N, M) = \sum_{j\geq 1} \mathbb{1}_{\{Y_j(\boldsymbol{X},N)\leq r_N\}} \mathbb{1}_{\{Y_j(\boldsymbol{X},N+M)\leq r_M\}},$$

namely the number of cells with sample frequency less or equal than r_N which have population frequency less or equal than r_M. A nonparametric estimator of τ_{r_N,r_M} and an upper bound for the corresponding NMSE can be derived along lines similar to those applied in this paper for τ_1. Regarding a lower bound on the NMSE, however, things get more challenging. Technically, the main difference would be in the approximation theory involved. Instead of finding the best (uniform) polynomial approximation to $x \mapsto \exp\{-Bx\}$ on some interval, we would have to find the best polynomial approximation to $x \mapsto q(x)\exp\{-Bx\}$ where q is some polynomial. As we are concerned with lower bounds, this turns out to be a much more challenging problem. The interest in τ_{r_N,r_M} is not only motivated in context of disclosure risk assessment, but also in the broad area of biosciences. Indeed, the discrete functional τ_{0,r_M} corresponds to the number of unseen rare species in additional unobservable samples, which is a natural refinement of the number of unseen species considered in [19]. Work on these problems is ongoing.

APPENDIX A: NONPARAMETRIC ESTIMATORS OF THE DISCLOSURE RISK: PROOFS

For the sake of simplifying notation, throughout this section we write τ_1 instead of $\tau_1(\boldsymbol{X}, N, M)$, $\hat{\tau}_1$ instead of $\hat{\tau}_1(\boldsymbol{X}(N), N)$, and $\hat{\tau}_1^L$ instead of $\hat{\tau}_1^L(\boldsymbol{X}(N), N)$.

A.1. Details for the determination of the estimator (2). First observe that, according to the definition of τ_1, we can write the following identities:

$$\mathbb{E}\big[Z_i(\boldsymbol{X}, N)\big] = \sum_{j\geq 1} \mathbb{P}\big(Y_j(\mathbf{X}, N) = i\big) = \sum_{j\geq 1} e^{-np_j} \frac{(np_j)^i}{i!}. \tag{A.1}$$

Then

$$\mathbb{E}[\tau_1] = \sum_{j\geq 1} \mathbb{P}\big(Y_j(\mathbf{X}, N) = 1\big)\mathbb{P}\big(Y_j(\mathbf{X}, N+M) - Y_j(\mathbf{X}, N) = 0\big) = \sum_{j\geq 1} np_j e^{-np_j} e^{-\lambda np_j},$$

and by a direct application of Taylor series expansion of the exponential function $e^{-\lambda np_j}$, for any $j \geq 1$, we can write the following expression:

$$\mathbb{E}[\tau_1] = \sum_{i\geq 0} \frac{(-1)^i \lambda^i}{i!} \sum_{j\geq 1} (np_j)^{i+1} e^{-np_j} = \sum_{i\geq 0} (-1)^i \lambda^i (i+1) \mathbb{E}\big[Z_{i+1}(\boldsymbol{X}, N)\big],$$

where the last equality follows from a direct application of the identity displayed in (A.1).

A.2. Empirical Bayes approach to determine (3). The estimator $\hat{\tau}_1$ admits a natural interpretation as a nonparametric empirical Bayes estimator in the sense of [22], that is, it is the posterior expectation of $\mathbb{E}[\tau_1]$ with respect to an empirical nonparametric prior distribution on the unknown p_j's. Specifically, note that $\mathbb{E}[\tau_1] = \sum_{j=1}^{+\infty} e^{-(\lambda+1)np_j} np_j$, and assume that the p_j's are independent and distributed according to the empirical cumulative distribution function $G(p)$ of $p_{i_1}, \ldots, p_{i_k}$, corresponding to the k distinct cells arising from the cross classification of the initial sample, namely $G(p) := k^{-1} \sum_{1\leq t\leq k} \mathbb{1}_{\{p_{i_t}\leq p\}}$. Consider a cell j containing x individuals out of the initial sample of size N, where $x \geq 0$, then from equation (9) of [22]

$$\varphi_n(x) := \frac{\int e^{-(\lambda+1)np} np e^{-np} \frac{(np)^x}{x!} G(\mathrm{d}p)}{\int e^{-np} \frac{(np)^x}{x!} G(\mathrm{d}p)} \tag{A.2}$$

is the Bayes estimator of the quantity $e^{-(\lambda+1)np_j} np_j$ appearing $\mathbb{E}[\tau_1]$, for a cell j which contains x individuals out of the initial sample of size N. Now, rewrite $\varphi_n(x)$ as

$$\begin{aligned}
\varphi_n(x) &= \frac{\int e^{-(\lambda+1)np} np e^{-np} \frac{(np)^x}{x!} G(\mathrm{d}p)}{\int e^{-np} \frac{(np)^x}{x!} G(\mathrm{d}p)} \\
&= \frac{\sum_{i\geq 0} \frac{(-(\lambda+1))^i}{i!x!} (x+i+1)! \int \frac{(np)^{x+i+1}}{(x+i+1)!} e^{-np} G(\mathrm{d}p)}{\int e^{-np} \frac{(np)^x}{x!} G(\mathrm{d}p)} \\
&= \frac{\sum_{i\geq 0} \frac{(-(\lambda+1))^i}{i!x!} (x+i+1)! \mathbb{E}[Z_{x+i+1}(\boldsymbol{X}, N)]}{\mathbb{E}[Z_x(\boldsymbol{X}, N)]}.
\end{aligned}$$

Then the nonparametric Bayes estimator of $\mathbb{E}[\tau_1]$ is obtained summing up over all the possible cross classification of the observed cells, where we replace $\mathbb{E}[Z_x(\boldsymbol{X}, N)]$ by their empirical counterparts $Z_x(\boldsymbol{X}, N)$. Specifically, we can write the following:

$$\begin{aligned}\hat{\tau}_1 &= \sum_{x\geq 0} Z_x(\boldsymbol{X}, N)\frac{\sum_{i\geq 0}\frac{(-(\lambda+1))^i}{i!x!}(x+i+1)!Z_{x+i+1}(\boldsymbol{X}, N)}{Z_x(\boldsymbol{X}, N)}\\ &= \sum_{i\geq 0}(i+1)Z_{i+1}(\boldsymbol{X}, N)\sum_{x=0}^{i}\frac{i!}{(i-x)!x!}\big(-(\lambda+1)\big)^{i-x}\\ &= \sum_{i\geq 0}(-1)^i\lambda^i(i+1)Z_{i+1}(\boldsymbol{X}, N),\end{aligned}$$

which coincides with the estimator (3) obtained by means of the identity displayed in (3).

A.3. Proof of Theorem 1. Because of the independence of the random variables $\{Y_j(\boldsymbol{X}, N)\}_{j\geq 1}$, we may write the variance $\mathrm{Var}(\tau_1 - \hat{\tau}_1)$ as follows:

$$\begin{aligned}&\mathrm{Var}(\tau_1 - \hat{\tau}_1)\\ &\quad= \sum_{j\geq 1}\mathrm{Var}\Bigg(\sum_{i\geq 0}(-1)^i(i+1)\lambda^i\mathbb{1}_{\{Y_j(\boldsymbol{X},N)=i+1\}} - \mathbb{1}_{\{Y_j(\boldsymbol{X},N)=1\}}\mathbb{1}_{\{Y_j(\boldsymbol{X},N+M)=1\}}\Bigg)\\ &\quad= \sum_{j\geq 1}\mathbb{E}\Bigg[\sum_{i\geq 1}a_i\mathbb{1}_{\{Y_j(\boldsymbol{X},N)=i+1\}} + \mathbb{1}_{\{Y_j(\boldsymbol{X},N)=1\}}\big(a_0 - \mathbb{1}_{\{Y_j(\boldsymbol{X},N+M)=1\}}\big)\Bigg]^2,\end{aligned}$$

where we have defined $a_i := (-1)^i(i+1)\lambda^i$. Now, observe that the events $\{(Y_j(\boldsymbol{X}, N) = i)\}_{i\geq 1}$ are all disjoint, hence the variance $\mathrm{Var}(\tau_1 - \hat{\tau}_1)$ may be rewritten as

$$\begin{aligned}&\sum_{j\geq 1}\mathbb{E}\Bigg[\sum_{i\geq 1}a_i^2\mathbb{1}_{\{Y_j(\boldsymbol{X},N)=i+1\}} + \mathbb{1}_{\{Y_j(\boldsymbol{X},N)=1\}}\big(a_0 - \mathbb{1}_{\{Y_j(\boldsymbol{X},N+M)=1\}}\big)^2\Bigg]\\ &\quad= \sum_{j\geq 1}\mathbb{E}\Bigg[\sum_{i\geq 0}a_i^2\mathbb{1}_{\{Y_j(\boldsymbol{X},N)=i+1\}} - \mathbb{1}_{\{Y_j(\boldsymbol{X},N)=1\}}\mathbb{1}_{\{Y_j(\boldsymbol{X},N+M)=1\}}\Bigg]\end{aligned}$$

observing that $a_0 = 1$. Thus, simple calculations show that we can bound $\mathrm{Var}(\tau_1 - \hat{\tau}_1)$ as

$$\begin{aligned}\mathrm{Var}(\tau_1 - \hat{\tau}_1) &\leq \max_{j\geq 0}|a_j|^2\mathbb{E}\big[Z_{\bar{1}}(\boldsymbol{X}, N)\big] - \sum_{j\geq 1}e^{-n(\lambda+1)p_j}np_j\\ &= \max_{i\geq 0}|a_i|^2\mathbb{E}\big[Z_{\bar{1}}(\boldsymbol{X}, N)\big] - \frac{1}{\lambda+1}\mathbb{E}\big[Z_1(\boldsymbol{X}, N+M)\big].\end{aligned}\tag{A.3}$$

It remains to show that the a_i's have a maximum for $\lambda < 1$, which is attained when $i = i^* := \lfloor(2\lambda - 1)/(1-\lambda)\rfloor \vee 0$. Hence the thesis follows by (A.3), since $\max_{i\geq 0}|a_i| = \Psi(\lambda)$.

A.4. Proof of Theorem 2. First, we focus on the determination of the bound (6), concerning the bias. Remember the definition of both $\hat{\tau}_1^L$ and τ_1 to write

$$\mathbb{E}\big[\hat{\tau}_1^L - \tau_1\big] = -\mathbb{E}\Bigg[\sum_{i\geq 0}(-1)^i(i+1)\lambda^i\mathbb{P}(L\leq i-1)Z_{i+1}(\boldsymbol{X}, N)\Bigg],$$

where we have observed that nonsmoothed estimator $\hat{\tau}_1$ is unbiased. It is now easy to see that

$$\begin{aligned}\mathbb{E}[\hat{\tau}_1^L - \tau_1] &= -\mathbb{E}\Big[\sum_{i\geq 0}(-1)^i(i+1)\lambda^i \mathbb{P}(L \leq i-1) Z_{i+1}(\boldsymbol{X}, N)\Big] \\ &= -\sum_{i\geq 1}\sum_{j\geq 1}(-1)^i(i+1)\lambda^i \mathbb{P}(L \leq i-1)\mathbb{P}\big(Y_j(\boldsymbol{X}, N) = i+1\big) \\ &= -\sum_{i\geq 1}\sum_{j\geq 1}(-1)^i(i+1)\lambda^i \mathbb{P}(L \leq i-1) e^{-np_j}\frac{(np_j)^{i+1}}{(i+1)!} \\ &= -\sum_{j\geq 1} e^{-np_j} np_j \sum_{i\geq 1}(-1)^i \frac{(\lambda np_j)^i}{i!}\mathbb{P}(L \leq i-1).\end{aligned} \tag{A.4}$$

Now we focus on the evaluation of the sum with respect to i. If we set $y := \lambda np_j$, then

$$\sum_{i\geq 1}\frac{(-y)^i}{i!}\mathbb{P}(L \leq i-1) = \sum_{i=1}^{+\infty}\frac{(-y)^i}{i!}\sum_{k=0}^{i-1}\mathbb{P}(L=k) = \sum_{k=0}^{+\infty}\mathbb{P}(L=k)\sum_{i=k+1}^{+\infty}\frac{(-y)^i}{i!}$$

and remembering the definition of the incomplete gamma function we obtain that

$$\begin{aligned}\sum_{i\geq 1}(-1)^i\frac{y^i}{i!}\mathbb{P}(L \leq i-1) &= \sum_{k=0}^{+\infty}\mathbb{P}(L=k)\frac{e^{-y}}{k!}\int_0^{-y}\tau^k e^{-\tau}\,\mathrm{d}\tau \\ &= -e^{-y}\int_0^y e^s\,\mathbb{E}_L\Big[\frac{(-s)^L}{L!}\Big]\mathrm{d}s.\end{aligned}$$

Putting the previous expression in (A.4) and observing that $y = \lambda np_j$, (6) immediately follows. Now, in order to bound the variance of the difference between τ_1 and its estimator $\hat{\tau}_1^L$, recall that $\{Y_j(\boldsymbol{X}, N)\}_{j\geq 1}$ are independent. Then

$$\begin{aligned}\mathrm{Var}(\hat{\tau}_1^L - \tau_1) &= \mathrm{Var}\Bigg(\sum_{i\geq 0}(-1)^i(i+1)\lambda^i Z_{i+1}(\boldsymbol{X}, N)\mathbb{P}(L \geq i) \\ &\qquad - \sum_{j=1}^{+\infty}\mathbb{1}_{\{Y_j(\boldsymbol{X},N)=1\}}\mathbb{1}_{\{Y_j(\boldsymbol{X},N+M)=1\}}\Bigg) \\ &= \sum_{j=1}^{+\infty}\mathrm{Var}\Bigg(\sum_{i=0}^{+\infty}a_i\mathbb{1}_{\{Y_j(\boldsymbol{X},N)=i+1\}} - \mathbb{1}_{\{Y_j(\boldsymbol{X},N)=1\}}\mathbb{1}_{\{Y_j(\boldsymbol{X},N+M)=1\}}\Bigg),\end{aligned}$$

having defined $a_i := (-1)^i(i+1)\lambda^i\mathbb{P}(L \geq i)$ for any $i \geq 0$. Therefore, we can write

$$\begin{aligned}&\mathrm{Var}(\hat{\tau}_1^L - \tau_1) \\ &\quad\leq \sum_{j=1}^{+\infty}\mathbb{E}\Bigg[\Bigg(\sum_{i=0}^{+\infty}a_i\mathbb{1}_{\{Y_j(\boldsymbol{X},N)=i+1\}} - \mathbb{1}_{\{Y_j(\boldsymbol{X},N)=1\}}\mathbb{1}_{\{Y_j(\boldsymbol{X},N+M)=1\}}\Bigg)^2\Bigg] \\ &\quad= \sum_{j=1}^{+\infty}\mathbb{E}\Bigg[\sum_{i=1}^{+\infty}a_i^2\mathbb{1}_{\{Y_j(\boldsymbol{X},N)=i+1\}} + \mathbb{1}_{\{Y_j(\boldsymbol{X},N)=1\}}\big(a_0 - \mathbb{1}_{\{Y_j(\boldsymbol{X},N+M)=1\}}\big)^2\Bigg],\end{aligned}$$

where we have used the incompatibility of the events $\{(Y_j(\boldsymbol{X}, N) = i)\}$ for different values of j. We can proceed with the upper bound for the variance as follows:

$$
\begin{aligned}
&\operatorname{Var}\bigl(\hat{\tau}_1^L - \tau_1\bigr) \\
&\quad= \sum_{j=1}^{+\infty} \mathbb{E}\Biggl[\sum_{i=0}^{+\infty} a_i^2 \mathbb{1}_{\{Y_j(\boldsymbol{X},N)=i+1\}} - \mathbb{1}_{\{Y_j(\boldsymbol{X},N)=1\}} \mathbb{1}_{\{Y_j(\boldsymbol{X},N+M)=1\}}\Biggr] \\
&\quad\le \max_{i\ge 0} |a_i|^2 \mathbb{E}\bigl[Z_{\bar{1}}(\boldsymbol{X}, N)\bigr] - \sum_{j=1}^{+\infty} \mathbb{E}\bigl[\mathbb{1}_{\{Y_j(\boldsymbol{X},N)=1\}} \mathbb{1}_{\{Y_j(\boldsymbol{X},N+M)=1\}}\bigr] \qquad \text{(A.5)} \\
&\quad= \max_{i\ge 0} |a_i|^2 \mathbb{E}\bigl[Z_{\bar{1}}(\boldsymbol{X}, N)\bigr] - \sum_{j=1}^{+\infty} e^{-\lambda n p_j} e^{-n p_j} n p_j \\
&\quad= \max_{i\ge 0} |a_i|^2 \mathbb{E}\bigl[Z_{\bar{1}}(\boldsymbol{X}, N)\bigr] - \frac{1}{\lambda+1} \mathbb{E}\bigl[Z_1(\boldsymbol{X}, N+M)\bigr].
\end{aligned}
$$

Now, let observe that we can estimate the maximum value of the $|a_i|$'s as follows:

$$
\begin{aligned}
\max_{i\ge 0} |a_i| &= \max_{i\ge 0} (i+1)\lambda^i \mathbb{P}(L \ge i) = \max_{i\ge 0} (i+1)\lambda^i \sum_{k=i}^{+\infty} \mathbb{P}(L = k) \\
&\le \max_{i\ge 0} \sum_{k=i}^{+\infty} (i+1)\lambda^i \mathbb{P}(L = k) \le \sum_{k=0}^{+\infty} (k+1)\lambda^k \mathbb{P}(L = k) \\
&= \mathbb{E}_L\bigl[(L+1)\lambda^L\bigr].
\end{aligned}
$$

Replacing $\max_{i\ge 0} |a_i|$ with $\mathbb{E}_L[(L+1)\lambda^L]$ in (A.5), the upper bound of $\operatorname{Var}(\hat{\tau}_1^L - \tau_1)$ becomes

$$
\operatorname{Var}\bigl(\hat{\tau}_1^L - \tau_1\bigr) \le \bigl(\mathbb{E}_L\bigl[(L+1)\lambda^L\bigr]\bigr)^2 \mathbb{E}\bigl[Z_{\bar{1}}(\boldsymbol{X}, N)\bigr] - \frac{\mathbb{E}[Z_1(\boldsymbol{X}, N+M)]}{\lambda+1}.
$$

The proof is completed by putting together the previous upper bound for the variance and the one for the bias (6), from which the bound on the MSE (7) easily follows.

A.5. Proof of Proposition 1. To prove (8), we use Theorem 2, bounding the two terms appearing in (7) separately. In order to obtain an estimate of first term on the right-hand side of (7), we note that for any $y > 0$ the following holds:

$$
\begin{aligned}
-e^{-y} \int_0^y e^s \mathbb{E}_L\biggl[\frac{(-s)^L}{L!}\biggr] \mathrm{d}s &= -e^{-y} \int_0^y e^s \sum_{k=0}^{+\infty} e^{-\beta} \frac{\beta^k}{k!} \frac{(-s)^k}{k!} \mathrm{d}s \\
&= -e^{-y-\beta} \int_0^y e^s \sum_{k=0}^{+\infty} \frac{(\beta s)^k (-1)^k}{\Gamma(k+1) k!} \mathrm{d}s.
\end{aligned}
$$

Recall that the Bessel polynomial (see [18]) is defined as $J_0(z) := \sum_{k=0}^{+\infty} \frac{(-1)^k z^{2k}}{2^{2k}\Gamma(k+1)k!}$, and that $|J_0(z)| \le 1$. Therefore, we obtain the following inequality:

$$
\biggl| -e^{-y} \int_0^y e^s \mathbb{E}_L\biggl[\frac{(-s)^L}{L!}\biggr] \mathrm{d}s \biggr| \le e^{-(y+\beta)} \int_0^y e^s \bigl|J_0(2\sqrt{s\beta})\bigr| \mathrm{d}s \le e^{-\beta}\bigl(1 - e^{-y}\bigr),
$$

which may be applied to bound the first term on the right-hand side of (7), with $y = \lambda n p_j$. Precisely,

$$\begin{aligned}
&\left| \sum_{j\geq 1} e^{-p_j n(\lambda+1)} p_j n \int_0^{\lambda n p_j} e^s \mathbb{E}_L\left[\frac{(-s)^L}{L!}\right] \mathrm{d}s \right| \\
&\leq \sum_{j\geq 1} e^{-np_j} np_j e^{-\beta}\left(1 - e^{-\lambda n p_j}\right) \leq e^{-\beta} \sum_{j=1}^{+\infty} e^{-np_j} np_j \\
&= e^{-\beta}\mathbb{E}\big[Z_1(\boldsymbol{X}, N)\big] \leq e^{-\beta}\mathbb{E}[N] = e^{-\beta} n.
\end{aligned} \tag{A.6}$$

In order to upper bound the other term on the right-hand side of (7), we observe that

$$\begin{aligned}
\mathbb{E}_L\big[(L+1)\lambda^L\big] &= \sum_{k=0}^{+\infty} e^{-\beta}\frac{\beta^k}{k!}\lambda^k(k+1) = e^{-\beta}\left(\sum_{k=1}^{+\infty}\frac{(\beta\lambda)^k}{(k-1)!} + \sum_{k=0}^{+\infty}\frac{(\beta\lambda)^k}{k!}\right) \\
&= e^{-\beta}\left(e^{\beta\lambda} + \beta\lambda e^{\beta\lambda}\right) = e^{\beta(\lambda-1)}(1+\beta\lambda),
\end{aligned}$$

hence we get

$$\begin{aligned}
&\left(\mathbb{E}_L\big[(L+1)\lambda^L\big]\right)^2 \mathbb{E}\big[Z_{\bar{1}}(\boldsymbol{X}, N)\big] - \frac{1}{\lambda+1}\mathbb{E}\big[Z_1(\boldsymbol{X}, N+M)\big] \\
&\leq n e^{2\beta(\lambda-1)}(1+\beta\lambda)^2.
\end{aligned} \tag{A.7}$$

Using (A.6) and (A.7), one can now estimate the MSE (7) in the Poisson case and (8) follows. Because of (8) the NMSE can be bounded from above by

$$\mathscr{E}_{n,\lambda}\big(\hat{\tau}_1^L\big) \leq e^{-2\beta} + \frac{e^{2\beta(\lambda-1)}(1+\beta\lambda)^2}{n}$$

using the exponential inequality $1 + x \leq e^x$ we get

$$\mathscr{E}_{n,\lambda}\big(\hat{\tau}_1^L\big) \leq e^{-2\beta} + \frac{e^{2\beta(2\lambda-1)}}{n}. \tag{A.8}$$

It is easy to show that the right-hand side of (A.8) is minimized when β equals $\frac{1}{4\lambda}\log(\frac{n}{2\lambda-1})$. Therefore, it is easy to observe that the inequality (A.8) becomes

$$\mathscr{E}_{n,\lambda}\big(\hat{\tau}_1^L\big) \leq \frac{1}{n^{1/(2\lambda)}} \cdot \frac{2\lambda}{(2\lambda-1)^{1-1/(2\lambda)}} \tag{A.9}$$

hence the second bound (9) follows provided that $A(\lambda) := \frac{2\lambda}{(2\lambda-1)^{1-1/(2\lambda)}}$. Now we can prove the "limit of predictability" in the Poisson case, indeed thanks to (9) we have

$$\mathscr{E}_{n,\lambda}\big(\hat{\tau}_1^L\big) \leq \frac{A}{n^{1/(2\lambda)}},$$

besides observe that $\frac{A}{n^{1/(2\lambda)}} \leq \delta$ is satisfied if and only if $\lambda \leq \frac{\log(n)}{2\log(A/\delta)} =: \lambda^*$. As a consequence the maximum value of λ for which the inequality $\mathscr{E}_{n,\lambda}(\hat{\tau}_1^L) \leq \delta$ is satisfied, is bigger or equal than λ^*, that is,

$$\max\big\{\lambda : \mathscr{E}_{n,\lambda}\big(\hat{\tau}_1^L\big) \leq \delta\big\} \geq \frac{\log(n)}{2\log(A/\delta)}.$$

Then the thesis follows by taking the limit of the previous inequality as $n \to +\infty$.

APPENDIX B: PROOFS RELATED TO THE LOWER BOUND

B.1. Proof of Lemma 1. First, it is obvious that

$$\mathscr{E}(\lambda, n) \le \inf_{\hat{\rho}} \sup_{P \in \mathscr{P}} n^{-2} \mathbb{E}_P^{n,\lambda}\big[\big(\tau_1(\boldsymbol{X}, N, M) - \hat{\rho}\big(\boldsymbol{Y}(\boldsymbol{X}, N)\big)\big)^2\big].$$

We now prove that the previous is indeed an inequality by deriving a lower bound that matches. Let $n > 0$ be fixed. By definition, for every $\varepsilon > 0$ there exists an estimator $\hat{\rho}_1$ such that

$$\begin{aligned}
\mathscr{E}(\lambda, n) &\ge \sup_{P \in \mathscr{P}} n^{-2} \mathbb{E}_P^{n,\lambda}\big[\big(\tau_1(\boldsymbol{X}, N, M) - \hat{\rho}_1\big(\boldsymbol{X}(N), N\big)\big)^2\big] - \varepsilon \\
&= \sup_{P \in \mathscr{P}} n^{-2} \mathbb{E}_P^{n,\lambda}\big[\mathbb{E}_P^{n,\lambda}\big[\big(\tau_1(\boldsymbol{X}, N, M) \\
&\qquad - \hat{\rho}_1\big(\boldsymbol{X}(N), N\big)\big)^2 \mid \boldsymbol{Y}(\boldsymbol{X}, N), \boldsymbol{Y}(\boldsymbol{X}, N + M)\big]\big] - \varepsilon \\
&\ge \sup_{P \in \mathscr{P}} n^{-2} \mathbb{E}_P^{n,\lambda}\big[\big(\tau_1(\boldsymbol{X}, N, M) - \mathbb{E}_P^{n,\lambda}\big[\hat{\rho}_1\big(\boldsymbol{X}(N), N\big) \mid \boldsymbol{Y}(\boldsymbol{X}, N)\big]\big)^2\big] - \varepsilon,
\end{aligned} \tag{B.1}$$

where the last line follows by Jensen's inequality and by observing that

$$\mathbb{E}_P^{n,\lambda}\big[\tau_1(\boldsymbol{X}, N, M) \mid \boldsymbol{Y}(\boldsymbol{X}, N), \boldsymbol{Y}(\boldsymbol{X}, N + M)\big] = \tau_1(\boldsymbol{X}, N, M) \quad \text{and}$$

$$\mathbb{E}_P^{n,\lambda}\big[\hat{\rho}_1\big(\boldsymbol{X}(N), N\big) \mid \boldsymbol{Y}(\boldsymbol{X}, N), \boldsymbol{Y}(\boldsymbol{X}, N + M)\big] = \mathbb{E}_P^{n,\lambda}\big[\hat{\rho}_1\big(\boldsymbol{X}(N), N\big) \mid \boldsymbol{Y}(\boldsymbol{X}, N)\big].$$

To see that the last equation is true, remark that $\boldsymbol{Y}(\boldsymbol{X}, N + M) - \boldsymbol{Y}(\boldsymbol{X}, N)$ is independent of $\boldsymbol{Y}(\boldsymbol{X}, N)$ and depends only on $(X_{N+1}, \ldots, X_{N+M})$. Now we claim that $\hat{\rho}_1$ can be chosen such that for any $k \in \mathbb{Z}_+$ and any permutation $\sigma_k(\boldsymbol{X}(k))$ of the data, it holds $\hat{\rho}_1(\boldsymbol{X}(k), k) = \hat{\rho}_1(\sigma_k(\boldsymbol{X}(k)), k)$. We delay the proof of the claim to later. Now assume the claim is true. Given k and $\boldsymbol{Y}(\boldsymbol{X}, k)$, we can construct the functional

$$G\big(\boldsymbol{Y}(\boldsymbol{X}, k), k\big) := (\underbrace{1, \ldots, 1}_{\times Y_1(\boldsymbol{X}, k)}, \underbrace{2, \ldots, 2}_{\times Y_2(\boldsymbol{X}, k)}, \ldots).$$

Since $\hat{\rho}_1$ is invariant under permutations of the data, we have for any $P \in \mathscr{P}$,

$$\begin{aligned}
&\mathbb{E}_P^{n,\lambda}\big[\hat{\rho}_1\big(\boldsymbol{X}(N), N\big) \mid \boldsymbol{Y}(\boldsymbol{X}, N)\big] \\
&\quad = \mathbb{E}_P^{n,\lambda}\big[\mathbb{E}_P^{n,\lambda}\big[\hat{\rho}_1\big(\boldsymbol{X}(N), N\big) \mid \boldsymbol{Y}(\boldsymbol{X}, N), N\big] \mid \boldsymbol{Y}(\boldsymbol{X}, N)\big] \\
&\quad = \mathbb{E}_P^{n,\lambda}\big[\mathbb{E}_P^{n,\lambda}\big[\hat{\rho}_1\big(G\big(\boldsymbol{Y}(\boldsymbol{X}, N), N\big), N\big) \mid \boldsymbol{Y}(\boldsymbol{X}, N), N\big] \mid \boldsymbol{Y}(\boldsymbol{X}, N)\big] \\
&\quad = \mathbb{E}_P^{n,\lambda}\big[\hat{\rho}_1\big(G\big(\boldsymbol{Y}(\boldsymbol{X}, N), N\big), N\big) \mid \boldsymbol{Y}(\boldsymbol{X}, N)\big] \\
&\quad = \hat{\rho}_1\big(G\big(\boldsymbol{Y}(\boldsymbol{X}, N), N\big), N\big).
\end{aligned}$$

The last line follows because $N = \sum_{j \ge 1} Y_j(\boldsymbol{X}, N)$, and hence N is completely determined by $\boldsymbol{Y}(\boldsymbol{X}, N)$. Therefore, we have proved that the conditional expected value of $\hat{\rho}_1(\boldsymbol{X}(N), N)$, given $\boldsymbol{Y}(\boldsymbol{X}, N)$ does not depend on P. Thus, (B.1) implies,

$$\begin{aligned}
\mathscr{E}(\lambda, n) &\ge \sup_{P \in \mathscr{P}} n^{-2} \mathbb{E}_P^{n,\lambda}\big[\big(\tau_1(\boldsymbol{X}, N, M) - \hat{\rho}_1\big(G\big(\boldsymbol{Y}(\boldsymbol{X}, N), N\big), N\big)\big)^2\big] - \varepsilon \\
&\ge \inf_{\hat{\rho}} \sup_{P \in \mathscr{P}} n^{-2} \mathbb{E}_P^{n,\lambda}\big[\big(\tau_1(\boldsymbol{X}, N, M) - \hat{\rho}\big(\boldsymbol{Y}(\boldsymbol{X}, N)\big)\big)^2\big] - \varepsilon.
\end{aligned}$$

Since the previous is true for all $\varepsilon > 0$, the conclusion follows.

We now prove the claim we have used in the previous argument, that is, that *$\hat{\rho}_1$ can be chosen such for any $k \in \mathbb{Z}_+$ and any permutation $\sigma_k(\boldsymbol{X}(k))$ of the data, it holds $\hat{\rho}_1(\boldsymbol{X}(k), k) =$*

$\hat{\rho}_1(\sigma_k(\boldsymbol{X}(k)), k)$. When $k = 0$, then the claim is trivial, hence we assume without loss of generality that $k \geq 1$. We will prove that for any estimator $\hat{\rho}_1$, there is a symmetric estimator $\hat{t}_1$ with a risk no more than the risk of $\hat{\rho}_1$. Let $\hat{\rho}_1$ be arbitrary. Construct $\hat{t}_1$ such that for any $k \in \mathbb{N}$,

$$\hat{t}_1(\boldsymbol{X}(k), k) := \frac{1}{|\{\sigma_k\}|} \sum_{\{\sigma_k\}} \hat{\rho}_1\big(\sigma_k(\boldsymbol{X}(k)), k\big).$$

Clearly, $\hat{t}_1$ has the desired invariance property under permutations. Moreover, by Jensen's inequality,

$$\begin{aligned}
&\mathbb{E}_P^{n,\lambda}\big[\big(\tau_1(\boldsymbol{X}, N, M) - \hat{t}_1\big(\boldsymbol{X}(N), N\big)\big)^2\big] \\
&\quad = \mathbb{E}_P^{n,\lambda}\bigg[\mathbb{E}_P^{n,\lambda}\bigg[\Big(\frac{1}{|\{\sigma_N\}|} \sum_{\{\sigma_N\}} \big(\tau_1(\boldsymbol{X}, N, M) - \hat{\rho}_1\big(\sigma_N\big(\boldsymbol{X}(N)\big), N\big)\big)^2 \,\Big|\, N\bigg]\bigg] \\
&\quad \leq \mathbb{E}_P^{n,\lambda}\bigg[\mathbb{E}_P^{n,\lambda}\bigg[\frac{1}{|\{\sigma_N\}|} \sum_{\{\sigma_N\}} \big(\tau_1(\boldsymbol{X}, N, M) - \hat{\rho}_1\big(\sigma_N\big(\boldsymbol{X}(N)\big), N\big)\big)^2 \,\Big|\, N\bigg]\bigg]
\end{aligned}$$

Now remark that for all $(k, k') \in \mathbb{Z}_+^2$ the map $\boldsymbol{X} \mapsto \tau_1(\boldsymbol{X}, k, k')$ is invariant under any permutations of the k first entries of $\boldsymbol{X}$. Moreover, $\boldsymbol{X}$ is an i.i.d. vector, then the last display implies that

$$\begin{aligned}
&\mathbb{E}_P^{n,\lambda}\big[\big(\tau_1(\boldsymbol{X}, N, M) - \hat{t}_1\big(\boldsymbol{X}(N), N\big)\big)^2\big] \\
&\quad \leq \mathbb{E}_P^{n,\lambda}\bigg[\mathbb{E}_P^{n,\lambda}\bigg[\frac{1}{|\{\sigma_N\}|} \sum_{\{\sigma_N\}} \big(\tau_1(\boldsymbol{X}, N, M) - \hat{\rho}_1\big(\boldsymbol{X}(N), N\big)\big)^2 \,\Big|\, N\bigg]\bigg] \\
&\quad = \mathbb{E}_P^{n,\lambda}\big[\big(\tau_1(\boldsymbol{X}, N, M) - \hat{\rho}_1\big(\boldsymbol{X}(N), N\big)\big)^2\big].
\end{aligned}$$

The conclusion follows by taking the supremum over $P \in \mathscr{P}$ both sides of the last display.

B.2. Proof of Lemma 2. In the whole proof, we drop the subscripts ξ whenever it is convenient.

Let $\sigma : [-1, 1] \to [\xi^{-1}, 1]$ be such that $\sigma(x) := (1 - \xi^{-1})(x + 1)/2 + \xi^{-1}$. Notice that σ is bijective. By translating and rescaling, we claim that $E_L(g, [\xi^{-1}, 1]) = E_L(g \circ \sigma, [-1, 1])$. To see that this is true, remark that for all $p \in \mathsf{P}_L[-1, 1]$ we have $\|g \circ \sigma - p\|_\infty = \|g - p \circ \sigma^{-1}\|_\infty \geq E_L(g, [\xi^{-1}, 1])$. This shows that $E_L(g \circ \sigma, [-1, 1]) \geq E_L(g, [\xi^{-1}, 1])$. The same steps using σ^{-1} show that $E_L(g \circ \sigma, [-1, 1]) \leq E_L(g, [\xi^{-1}, 1])$. Hence $E_L(g, [\xi^{-1}, 1]) = E_L(g \circ \sigma, [-1, 1])$.

For the sake of simplicity, we let $C := B(1 - \xi^{-1})$ and $\gamma_C : [-1, 1] \to \mathbb{R}_+$ is defined by $\gamma_C(x) = \exp\{-C(x + 1)\}$. From the discussion in the previous paragraph, we have indeed reduced the problem to finding $E_L(\gamma_C, [-1, 1])$. This is because

$$\begin{aligned}
E_L\big(g, [\xi^{-1}, 1]\big) = E_L\big(g \circ \sigma, [-1, 1]\big) &= \exp\big\{-2B\xi^{-1}\big\} E_L\big(\gamma_C, [-1, 1]\big) \\
&= e\big(1 + o(1)\big) E_L\big(\gamma_C, [-1, 1]\big).
\end{aligned}$$

To find a lower bound on $E_L(\gamma_C, [-1, 1])$, we will exploit the well-known relationship between uniform approximation on the interval by polynomials and uniform approximation of periodic even functions by trigonometric polynomials. We write $\mathsf{CE}[-1, 1]$ the space of continuous and even functions on $[-1, 1]$, and for any $L \in \mathbb{Z}_+$ we let $\mathsf{TP}_L[-1, 1]$ denote the set of even trigonometric polynomials of degree at most L, that is, $\mathsf{TP}_L[-1, 1]$ is

$$\bigg\{T \in \mathsf{CE}[-1, 1] : T(x) = \sum_{k=0}^{L} a_k \cos(\pi k x), a_k \in \mathbb{R}, x \in [-1, 1]\bigg\}.$$

We furthermore define the periodization operator $P : \mathsf{C}[-1,1] \to \mathsf{CE}[-1,1]$ such that $Pf(\theta) = f(\cos(\pi\theta))$ for all $f \in \mathsf{C}[-1,1]$ and all $\theta \in [-1,1]$. Then it is well known (see, for instance, the Theorem 14.8.1 in [6]) that

$$E_L(\gamma_C, [-1,1]) = \inf\{\|P\gamma_C - T\|_\infty : T \in \mathsf{TP}_L[-1,1]\}. \tag{B.2}$$

We will now bound the right-hand side of (B.2) by a technique inspired from [16], which works as well for our setting. For any $K \in \mathbb{N}$, we define the trigonometric polynomial $T_K : [-1,1] \to \mathbb{C}$ such that

$$T_K(\theta) := e^{i\pi(L+1)\theta}\left\{\sum_{k=0}^{K-1} e^{i2\pi k\theta}\right\}^2.$$

Then, by orthogonality of the trigonometric polynomials, we have that

$$\int_{-1}^{-1} |T_K(\theta)|\,\mathrm{d}\theta = \sum_{j=0}^{K-1}\sum_{k=0}^{K-1}\int_{-1}^{1} e^{i2\pi(j-k)\theta}\,\mathrm{d}\theta = 2K. \tag{B.3}$$

By definition, for every $\varepsilon > 0$ we can find a $Q \in \mathsf{TP}_L[-1,1]$ such that $\|P\gamma_C - Q\|_\infty \le E_L(\gamma_C, [-1,1]) + \varepsilon$. Choose such Q, and remark that (B.3) implies

$$\begin{aligned}\left|\int_{-1}^{1}\big(P\gamma_C(\theta) - Q(\theta)\big)T_K(\theta)\,\mathrm{d}\theta\right| &\le \|P\gamma_C - Q\|_\infty \int_{-1}^{1}|T_K(\theta)|\,\mathrm{d}\theta \\ &\le 2K\{E_L(\gamma_C, [-1,1]) + \varepsilon\}.\end{aligned}$$

On the other hand remark that Q is a trigonometric polynomial of degree at most L, while T_K is a trigonometric polynomial of degree strictly greater than L. Therefore, Q is orthogonal to T_K. Moreover, the last display is true for all $\varepsilon > 0$ and for all $K \in \mathbb{N}$, thus it must be the case that

$$E_L(\gamma_C, [-1,1]) \ge \max_{K\in\mathbb{N}} \frac{1}{2K}\left|\int_{-1}^{1} P\gamma_C(\theta)T_K(\theta)\,\mathrm{d}\theta\right|. \tag{B.4}$$

Interestingly, we can compute the previous integral. Namely,

$$\begin{aligned}\int_{-1}^{1} P\gamma_C(\theta)T_K(\theta)\,\mathrm{d}\theta &= \sum_{j=0}^{K-1}\sum_{k=0}^{K-1}\int_{-1}^{1}\gamma_C\big(\cos(\pi\theta)\big)e^{i\pi\theta(L+1+2j+2k)}\,\mathrm{d}\theta \\ &= 2(-1)^{L+1}\sum_{j=0}^{K-1}\sum_{k=0}^{K-1} e^{-C} I_{L+1+2j+2k}(C),\end{aligned}$$

where $I_\nu(z) := \frac{1}{\pi}\int_0^\pi e^{z\cos(t)}\cos(\nu t)\,\mathrm{d}t$ is the modified Bessel function (see [18], p. 248); in particular [18], formula 10.32.3. More precisely, from the above considerations and the fact that the modified Bessel functions are nonnegative, we deduce that

$$\left|\int_{-1}^{1} P\gamma_C(\theta)T_K(\theta)\,\mathrm{d}\theta\right| = 2\sum_{j=0}^{K-1}\sum_{k=0}^{K-1} e^{-C} I_{L+1+2j+2k}(C).$$

Soni [27] proved that $I_{k+1}(z) \le I_k(z)$ for all $k \in \mathbb{N}$ and all $z > 0$. Hence, we obtain from the last display and (B.4) the bound

$$E_L(\gamma_C, [-1,1]) \ge \max_{K\in\mathbb{N}} K e^{-C} I_{L+4K}(C). \tag{B.5}$$

In the next lemma, We obtain a bound on the modified Bessel function $z \mapsto I_k(z)$ which remains tighter than the classical bound derived in [12] when $z \ge k$. The proof of the lemma is to be found in Section B.3.

LEMMA B.1. *Assume $k \in \mathbb{N}$ and assume that $C > 8\sqrt{1+(k/C)^2}$. Then*

$$e^{-C} I_k(C) > \frac{\exp\{-C\varphi(k/C)\}}{2e^4(1+(k/C)^2)^{1/4}\sqrt{C}}.$$

For $\alpha, \beta \in \mathbb{R}$ to be chosen accordingly, we define $K_* := \alpha\sqrt{C}$ if $L < \sqrt{C}$, or $K_* := \beta C/L$ if $L \geq \sqrt{C}$. In view of (B.5), it is clear that $E_L(\gamma_C, [-1,1]) \geq K_* e^{-C} I_{L+4K_*}(C)$. Consider now the case where $L < \sqrt{C}$, then

$$0 \leq \frac{L+4K_*}{C} = \frac{L+\alpha\sqrt{C}}{C} < \frac{\alpha+1}{\sqrt{C}}.$$

Thus, $C\varphi((L+4K_*)/C) = O(1)$ as $C \to \infty$, $(L+4K_*)/C \to 0$ as $C \to \infty$, and $C > 8\sqrt{1+(L+4K_*)^2/C^2}$ when C gets large enough. We then obtain from Lemma B.1 that in this case,

$$E_L(\gamma_C, [-1,1]) > \frac{\alpha\sqrt{C}(1+o(1))\exp\{-C\varphi((L+4K_*)/C)\}}{2e^2\sqrt{C}} \gtrsim 1,$$

at least for C large enough. We now consider the case $L \geq \sqrt{C}$. In this case, we have

$$0 \leq \frac{L+4K_*}{C} = \frac{L+\beta C/L}{C} = \frac{L}{C} + \frac{\beta}{L} \leq \frac{L}{C} + \frac{\beta}{\sqrt{C}}.$$

Because by assumption there is a constant $\zeta > 0$ such that $L \leq \zeta C$, then $(L+4K_*)/C \leq \zeta + o(1)$ as $C \to \infty$, and thus we have $C > 8\sqrt{1+(L+4K_*)^2/C^2}$ when C is large enough. Then we can apply Lemma B.1 to find that as $C \to \infty$,

$$E_L(\gamma_C, [-1,1]) > \frac{(\beta C/L)(1+o(1))\exp\{-C\varphi((L+4K_*)/C)\}}{4e^2\sqrt{C}(1+(L/C)^2)^{1/4}}$$

$$\gtrsim \sqrt{\frac{C}{L^2}} \exp\left\{-C\varphi\left(\frac{L}{C}+\frac{\beta}{L}\right)\right\},$$

at least for C large enough. Further, it can be seen that $|\varphi'(x)| \leq |x|$ (see, for instance, Section S6.3 in the Supplementary Material [3]). Then, by Taylor expansion,

$$\varphi\left(\frac{L}{C}+\frac{\beta}{L}\right) \leq \varphi\left(\frac{L}{C}\right) + \left(\frac{L}{C}+\frac{\beta}{L}\right)\frac{\beta}{L},$$

and thus, $C\varphi(L/C+\beta/L) \leq C\varphi(L/C) + \beta(1+C\beta/L^2) \leq C\varphi(L/C) + \beta(1+\beta)$. It follows

$$E_L(\gamma_C, [-1,1]) \gtrsim \sqrt{\frac{C}{L^2}} \exp\left\{-C\varphi\left(\frac{L}{C}\right)\right\}.$$

With similar arguments, $C\varphi(L/C) = (\xi/2)\varphi(2L/\xi) + O(1)$ as $\xi \to \infty$.

B.3. Proof of Lemma B.1. The proof relies on the well known series representation of the modified Bessel function (see [18], formula 10.25.2), namely we have whenever $k \in \mathbb{N}$,

$$I_k(z) = \sum_{p=0}^{\infty} \frac{1}{p!(p+k)!}\left(\frac{z}{2}\right)^{2p+k}. \tag{B.6}$$

Conveniently, all the terms in the summation are nonnegative, which we will exploit to get our lower bound. By Stirling's formula, when $k \geq 1$, for any $p \geq 0$,

$$(p+k)! \leq e\sqrt{(p+k)}\exp\{-(p+k)+(p+k)\log(p+k)\},$$

and for any $p \geq 1$, we have $p! \leq e\sqrt{p}\exp\{-p + p\log p\}$. For convenience, let define the functions $\phi_{z,k} : \mathbb{R}_+^* \to \mathbb{R}_+$, such that for any $x, z \in \mathbb{R}_+^*$ and any $k \in \mathbb{N}$,

$$\phi_{z,k}(x) := -z + 2x + k - x\log x - (x+k)\log(x+k) + (2x+k)\log(z/2).$$

Hence, because each term in the series expansion of (B.6) is nonnegative, we get the estimate

$$e^{-z} I_k(z) \geq e^{-z} \sum_{p\geq 1} \frac{1}{p!(p+k)!}\left(\frac{z}{2}\right)^{2p+k} \geq \frac{1}{e^2}\sum_{p\geq 1}\frac{\exp\{\phi_{z,k}(p)\}}{\sqrt{p(p+k)}}. \tag{B.7}$$

Notice that

$$\phi'_{z,k}(x) = -\log(x) - \log(x+k) + 2\log(z/2), \qquad \phi''_{z,k}(x) = -\frac{1}{x} - \frac{1}{x+k}.$$

Thus, $\phi_{z,k}$ admits a unique nonnegative extremum at x_0 solution to $x_0(x_0+k) = z^2/4$, that is,

$$x_0 = \frac{-k + \sqrt{k^2+z^2}}{2} \quad \text{and} \quad \phi''_{z,k}(x_0) = -\frac{4}{z}\sqrt{1+(k/z)^2} < 0.$$

Henceforth x_0 is indeed the unique maximum of the function $\phi_{z,k}$ on $\mathbb{R}_+$. We let p_0 smallest integer larger than x_0. Then $p_0 \geq 1$ and we have, by Taylor expansion that for any $p \geq p_0$ there is a $\bar{p} \in (x_0, p)$

$$\begin{aligned}\phi_{z,k}(p) &= \phi_{z,k}(x_0) + \phi'_{z,k}(x_0)(p-x_0) + \frac{1}{2}\phi''_{z,k}(\bar{p})(p-x_0)^2 \\ &= \phi_{z,k}(x_0) + \frac{1}{2}\phi''_{z,k}(\bar{p})(p-x_0)^2.\end{aligned}$$

Remark that, because $\bar{p} \geq x_0$,

$$\phi''_{z,k}(\bar{p}) = -\frac{1}{\bar{p}} - \frac{1}{\bar{p}+k} \geq -\frac{1}{x_0} - \frac{1}{x_0+k} = -\frac{4}{z}\sqrt{1+(k/z)^2}.$$

Then, for any $p \geq p_0$,

$$\phi_{z,k}(p) \geq \phi_{z,k}(x_0) + \frac{1}{2}\phi''_{z,k}(x_0)(p-x_0)^2 = \phi_{z,k}(x_0) - \frac{2\sqrt{1+(k/z)^2}}{b}(p-x_0)^2.$$

Therefore,

$$e^{-z} I_k(z) \geq \frac{\exp\{\phi_{z,k}(x_0)\}}{e^2}\sum_{p\geq p_0}\frac{\exp\{\phi''_{z,k}(x_0)(p-x_0)^2/2\}}{\sqrt{p(p+k)}}.$$

Let p_1 be the largest integer such that $-\phi''_{z,k}(x_0)(p_1-x_0)^2 \leq 2$. Remark that whenever $z > 2(1+(k/z)^2)^{1/2}$, we have $p_1 \geq x_0 + 1$, which is always the case in the conditions of the lemma. Because the summand is the previous is monotonically decreasing for $p \geq p_0$, we get the bound

$$e^{-z} I_k(z) \geq \frac{\exp\{\phi_{z,k}(x_0)\}}{e^4}\frac{(p_1-p_0)}{\sqrt{p_1(p_1+k)}} \geq \frac{\exp\{\phi_{z,k}(x_0)\}}{e^4}\frac{(p_1-x_0)-1}{\sqrt{p_1(p_1+k)}}.$$

But, by the definition of p_1, we have that $p_1 + 1 - x_0 > \sqrt{2/(-\phi''_{z,k}(x_0))}$. Therefore, whenever $z > 8(1+(k/z)^2)^{1/2}$, by the definition of $\phi''_{z,k}(x_0)$,

$$\begin{aligned}e^{-z} I_k(z) &\geq \frac{\exp\{\phi_{z,k}(x_0)\}}{e^4\sqrt{-\phi''_{z,k}(x_0)p_1(p_1+k)}}\{\sqrt{2} - 2\sqrt{-\phi''_{z,k}(x_0)}\} \\ &\geq \frac{\sqrt{2}\exp\{\phi_{z,k}(x_0)\}}{2e^4\sqrt{-\phi''_{z,k}(x_0)p_1(p_1+k)}}.\end{aligned}$$

Also,

$$\begin{aligned} p_1(p_1+k) &= x_0(x_0+k) + \left(p_1^2 - x_0^2\right) + (p_1 - x_0)k \\ &= x_0(x_0+k) + (p_1 - x_0)(p_1 + x_0 + k) \\ &= x_0(x_0+k) + (p_1 - x_0)^2 + (p_1 - x_0)(2x_0 + k). \end{aligned}$$

But we have that $x_0(x_0+k) = z^2/4$, $(p_1 - x_0)^2 \le -2/\phi''_{z,k}(x_0)$, and $2x_0 + k = z\sqrt{1+(k/z)^2}$. Thus,

$$\begin{aligned} p_1(p_1+k) &\le \frac{z^2}{4} + \frac{2}{-\phi''_{z,k}(x_0)} + \sqrt{\frac{2(1+(k/z)^2)}{-\phi''_{z,k}(x_0)}}z \\ &= \frac{z^2}{4} + \frac{z}{2\sqrt{1+(k/z)^2}} + \frac{z^{3/2}}{\sqrt{2}}\left[1+(k/z)^2\right]^{1/4} \\ &= \frac{z^2}{4}\left\{1 + \frac{z^{-1/2}[1+(k/z)^2]^{1/4}}{\sqrt{2}} + \frac{z^{-1}}{2\sqrt{1+(k/z)^2}}\right\}. \end{aligned}$$

Therefore, whenever $z > 8(1+(k/z)^2)^{1/2}$,

$$p_1(p_1+k) \le \frac{z^2}{4}\left\{1 + \frac{1}{4} + \frac{1}{16}\right\} \le \frac{21}{64}z^2 < \frac{z^2}{2}.$$

Hence,

$$e^{-z}I_k(z) > \frac{\exp\{\phi_{z,k}(x_0)\}}{e^4\sqrt{-\phi''_{z,k}(x_0)z}} = \frac{\exp\{\phi_{z,k}(x_0)\}}{2e^4(1+(k/z)^2)^{1/4}\sqrt{z}}.$$

After some algebra, we find that

$$\begin{aligned} \phi_{z,k}(x_0) &= -z + z\sqrt{1+(k/z)^2} \\ &\quad - (z/2)\left\{-(k/z) + \sqrt{1+(k/z)^2}\right\}\log\left\{-(k/z) + \sqrt{1+(k/z)^2}\right\} \\ &\quad - (z/2)\left\{(k/z) + \sqrt{1+(k/z)^2}\right\}\log\left\{(k/z) + \sqrt{1+(k/z)^2}\right\} \\ &= -z + z\sqrt{1+(k/z)^2} - z\cdot(k/z)\,\mathrm{arcsinh}(k/z) = -z\varphi(k/z). \end{aligned}$$

Acknowledgments. F. Camerlenghi also affiliated to Collegio Carlo Alberto (Turin, Italy) and BIDSA at Bocconi University (Milan, Italy).

S. Favaro also affiliated to Collegio Carlo Alberto (Turin, Italy) and IMATI-CNR "Enrico Magenes" (Milan, Italy).

The authors are grateful to an Associate Editor and two anonymous referees for all their comments, corrections and suggestions which remarkably improved the paper.

Federico Camerlenghi and Stefano Favaro received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement No 817257. Federico Camerlenghi and Stefano Favaro gratefully acknowledge the financial support from the Italian Ministry of Education, University and Research (MIUR), "Dipartimenti di Eccellenza" grant 2018-2022.

SUPPLEMENTARY MATERIAL

Supplement to "Optimal disclosure risk assessment" (DOI: 10.1214/20-AOS1975SUPP; .pdf). Supplementary S2 contains the complete proof of the minimax lower bound given in Theorem 3. Supplementary S3 contains an illustration on synthetic data for the estimators of Section 2, and their comparison with estimators from the existing literature. Supplementary S4 contains the proof that the lower bound on $E_L(g_\xi, [\xi^{-1}, 1])$ in Lemma 2 is sharp (up to constants). Supplementary S5 and S6 contain the proofs of the auxiliary results, respectively for the minimax lower bound and the tightness of the lower bound on $E_L(g_\xi, [\xi^{-1}, 1])$.

REFERENCES

[1] BEN-HAMOU, A., BOUCHERON, S. and OHANNESSIAN, M. I. (2017). Concentration inequalities in the infinite urn scheme for occupancy counts and the missing mass, with applications. *Bernoulli* **23** 249–287. MR3556773 https://doi.org/10.3150/15-BEJ743

[2] BETHLEHEM, J. G., KELLER, W. J. and PANNEKOEK, J. (1990). Disclosure control of microdata. *J. Amer. Statist. Assoc.* **85** 38–45.

[3] CAMERLENGHI, F., FAVARO, S., NAULET, Z. and PANERO, F. (2021). Supplement to "Optimal disclosure risk assessment." https://doi.org/10.1214/20-AOS1975SUPP

[4] CAROTA, C., FILIPPONE, M., LEOMBRUNI, R. and POLETTINI, S. (2015). Bayesian nonparametric disclosure risk estimation via mixed effects log-linear models. *Ann. Appl. Stat.* **9** 525–546. MR3341126 https://doi.org/10.1214/15-AOAS807

[5] CAROTA, C., FILIPPONE, M. and POLETTINI, S. (2018). Assessing Bayesian nonparametric log-linear models: An application to disclosure risk estimation. Preprint. Available at arXiv:1801.05244.

[6] DAVIDSON, K. R. and DONSIG, A. P. (2010). *Real Analysis and Applications*: *Theory in Practice*. *Undergraduate Texts in Mathematics*. Springer, New York. MR2568574 https://doi.org/10.1007/978-0-387-98098-0

[7] EFRON, B. and MORRIS, C. (1973). Stein's estimation rule and its competitors—An empirical Bayes approach. *J. Amer. Statist. Assoc.* **68** 117–130. MR0388597

[8] EFRON, B. and THISTED, R. (1976). Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika* **63** 435–447.

[9] FIENBERG, S. E. and MAKOV, U. E. (1998). Condentiality, uniqueness, and disclosure limitation for categorical data. *J. Off. Stat.* **14** 385–397.

[10] GOOD, I. J. and TOULMIN, G. H. (1956). The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika* **43** 45–63. MR0077039 https://doi.org/10.1093/biomet/43.1-2.45

[11] JIAO, J., VENKAT, K., HAN, Y. and WEISSMAN, T. (2015). Minimax estimation of functionals of discrete distributions. *IEEE Trans. Inf. Theory* **61** 2835–2885. MR3342309 https://doi.org/10.1109/TIT.2015.2412945

[12] LUKE, Y. L. (1972). Inequalities for generalized hypergeometric functions. *J. Approx. Theory* **5** 41–65. MR0350082 https://doi.org/10.1016/0021-9045(72)90028-7

[13] MANRIQUE-VALLIER, D. and REITER, J. P. (2012). Estimating identification disclosure risk using mixed membership models. *J. Amer. Statist. Assoc.* **107** 1385–1394. MR3036402 https://doi.org/10.1080/01621459.2012.710508

[14] MANRIQUE-VALLIER, D. and REITER, J. P. (2014). Bayesian estimation of discrete multivariate latent structure models with structural zeros. *J. Comput. Graph. Statist.* **23** 1061–1079. MR3270711 https://doi.org/10.1080/10618600.2013.844700

[15] MOSSEL, E. and OHANNESSIAN, M. I. (2019). On the impossibility of learning the missing mass. *Entropy* **21** Art. ID 28.

[16] NEWMAN, D. J. and RIVLIN, T. J. (1976). Approximation of monomials by lower degree polynomials. *Aequationes Math.* **14** 451–455. MR0410181 https://doi.org/10.1007/BF01835995

[17] OHANNESSIAN, M. I. and DAHLEH, M. A. (2012). Rare probability estimation under regularly varying heavy tails. *J. Mach. Learn. Res.* **23** 1–24.

[18] OLVER, F. W. J., LOZIER, D. W., BOISVERT, R. F. and CLARK, C. W. (2010). *NIST Handbook of Mathematical Functions*. Cambridge Univ. Press, Cambridge.

[19] ORLITSKY, A., SURESH, A. T. and WU, Y. (2016). Optimal prediction of the number of unseen species. *Proc. Natl. Acad. Sci. USA* **113** 13283–13288. MR3582444 https://doi.org/10.1073/pnas.1607774113

[20] REITER, J. P. (2005). Estimating risks of identification disclosure in microdata. *J. Amer. Statist. Assoc.* **100** 1103–1112. MR2236926 https://doi.org/10.1198/016214505000000619

[21] RINOTT, Y. and SHLOMO, N. (2006). A generalized negative binomial smoothing model for sample disclosure risk estimation. In *Privacy in Statistical Databases*. *Lecture Notes in Computer Science*. Springer, Berlin.

[22] ROBBINS, H. (1956). An empirical Bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1954–1955, *Vol. I* 157–163. Univ. California Press, Berkeley, CA. MR0084919

[23] SAMUELS, S. M. (1998). A Bayesian, species-sampling-inspired approach to the uniques problem in microdata disclosure risk assessment. *J. Off. Stat.* **14** 373–383.

[24] SKINNER, C., MARSH, C., OPENSHAW, S. and WYMER, C. (1994). Disclosure control for census microdata. *J. Off. Stat.* **10** 31–51.

[25] SKINNER, C. and SHLOMO, N. (2008). Assessing identification risk in survey microdata using log-linear models. *J. Amer. Statist. Assoc.* **103** 989–1001. MR2462887 https://doi.org/10.1198/016214507000001328

[26] SKINNER, C. J. and ELLIOT, M. J. (2002). A measure of disclosure risk for microdata. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **64** 855–867. MR1979391 https://doi.org/10.1111/1467-9868.00365

[27] SONI, R. P. (1965). On an inequality for modified Bessel functions. *J. Math. Phys.* **44** 406–407. MR0185164

[28] TSYBAKOV, A. B. (2009). *Introduction to Nonparametric Estimation*. *Springer Series in Statistics*. Springer, New York. MR2724359 https://doi.org/10.1007/b13794

[29] VALIANT, P. and VALIANT, G. (2013). Estimating the unseen: Improved estimators for entropy and other properties. *Adv. Neural Inf. Process. Syst.* **27** 2157–2165.

[30] WILLENBORG, L. and DE WAAL, T. (2001). *Elements of Statistical Disclosure Control*. *Lecture Notes in Statistics* **155**. Springer, New York. MR1866909 https://doi.org/10.1007/978-1-4613-0121-9

[31] WU, Y. and YANG, P. (2016). Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Trans. Inf. Theory* **62** 3702–3720. MR3506758 https://doi.org/10.1109/TIT.2016.2548468

[32] WU, Y. and YANG, P. (2019). Chebyshev polynomials, moment matching, and optimal estimation of the unseen. *Ann. Statist.* **47** 857–883. MR3909953 https://doi.org/10.1214/17-AOS1665

5.1 Supplementary material

SUPPLEMENTARY MATERIAL FOR "OPTIMAL DISCLOSURE RISK ASSESSMENT"

BY F. CAMERLENGHI[‡,*], S. FAVARO[§,†], Z. NAULET[¶] AND F. PANERO[‖]

University of Milano - Bicocca[‡], University of Torino[§], Université Paris-Saclay[¶] and University of Oxford[‖]

S1. Organization of the document. This document is the companion paper to the article *Optimal Disclosure Risk Assessment*, by the same authors. It complements the result of the main paper in the following way:

- In Section S2, we give the complete proof of the minimax lower bound given in Theorem 3 of the main document, with all details.
- In Section S3, we present an illustration on synthetic data of the estimators introduced in Section 2. We compare our estimator with various estimators from the existing literature.
- In Section S4, we demonstrate that the lower bound $E_L(g_\xi, [\xi^{-1}, 1])$ derived in Lemma 2 of the main document is sharp (up to constants). The proof is constructive and exhibits that Chebychev polynomials achieve the bound.
- Finally, Sections S5 and S6 contain the proofs of the auxiliary results, respectively for the minimax lower bound and the tightness of the lower bound on $E_L(g_\xi, [\xi^{-1}, 1])$.

S2. Complete proof of the minimax lower bound. This section is devoted to the complete proof of the minimax lower bound stated in the main document, that is Theorem 3. Unless specified otherwise, the notations and conventions are the same as in the main document. We recall that the minimax risk is defined as

$$\mathscr{E}(\lambda, n) \coloneqq \inf_{\hat{\rho}_1} \sup_{P \in \mathscr{P}} n^{-2} \mathbb{E}_P^{n,\lambda}[(\hat{\rho}_1(\boldsymbol{X}(N), N) - \tau_1(\boldsymbol{X}, N, M))^2], \tag{S1}$$

where the infinimum is taken over all estimators $\hat{\rho}_1$. To obtain a lower bound on the last display, we adapt the reduction scheme of [16, 15] which is based on the method of the *two fuzzy hypotheses* [14]. More precisely, the proof consists on the following steps.

[*] Also affiliated to Collegio Carlo Alberto (Torino, Italy) and BIDSA at Bocconi University (Milan, Italy).

[†] Also affiliated to IMATI-CNR "Enrico Magenes" (Milan, Italy).

Step 1. The very first step is to use Lemma 1 in the main document. We recall that Lemma 1 shows that the infinimum in equation (S1) can be restricted over estimators depending only on $(\boldsymbol{X}(N), N)$ through $\boldsymbol{Y}(\boldsymbol{X}, N)$. The details for this step are in the main document and omitted here. We recall the result

$$\mathscr{E}(\lambda, n) = \inf_{\hat{\rho}} \sup_{P \in \mathscr{P}} n^{-2} \mathbb{E}_P^{n,\lambda}[(\tau_1(\boldsymbol{X}, N, M) - \hat{\rho}(\boldsymbol{Y}(\boldsymbol{X}, N)))^2]. \tag{S2}$$

Step 2. The rhs of equation (S2) does not look like a classical minimax bound because $\tau_1(\boldsymbol{X}, N, M)$ is a random variable and not a function of $P \in \mathscr{P}$ (though its distribution is). In order to reduce the problem to a classical minimax problem, we show that τ_1 is sufficiently concentrated around its expectation so that $\tau_1(\boldsymbol{X}, N, M)$ can be traded (asymptotically as $n \to \infty$) for $\bar{\tau}_1(P, n, \lambda) := \mathbb{E}_P^{n,\lambda}[\tau_1(\boldsymbol{X}, N, M)]$ under the model P. This is made formal in the next proposition, proved in Section S5.1.

PROPOSITION S1. *Let $\boldsymbol{Y}_N$ denote the random variable $(\boldsymbol{X}, N) \mapsto \boldsymbol{Y}(\boldsymbol{X}, N)$. Then for any $\lambda, n > 0$ the following is true,*

$$\mathscr{E}(\lambda, n) \geq \frac{1}{2} \inf_{\hat{\rho}} \sup_{P \in \mathscr{P}} n^{-2} \mathbb{E}_P^{n,\lambda}[(\bar{\tau}_1(P, n, \lambda) - \hat{\rho}(\boldsymbol{Y}_N))^2] - n^{-1}. \tag{S3}$$

Remark that we dropped-out the superscript λ in $\mathbb{E}_P^{n,\lambda}$ in Proposition S1 as the argument in the expectation is independent of M, and thus its distribution depends on λ only through $\bar{\tau}_1(P, n, \lambda)$.

Step 3. The reduction scheme of [16, 15] involves the construction of (fuzzy) hypotheses that are not probability distributions, but only quasi probability distributions. Namely, to use their reduction scheme, we need to show that trading $\mathscr{P}$ for a suitable set of quasi probability distributions $\mathscr{P}'$ in equation (S3) does not affect the bound too much.

For $S \in \mathbb{N}$, $\xi, \delta > 0$ to be chosen accordingly at the end of the day, we define $\mathscr{P}'$ as

$$\mathscr{P}' := \left\{ \textstyle\sum_{k=1}^S p_k \delta_k : \; p_k \in [0, \xi S^{-1}], \; |\sum_{k=1}^S p_k - 1| \leq \delta \right\}. \tag{S4}$$

Here and after, under $\mathbb{P}_P^{n,\lambda}$ with $n > 0$ and $P \in \mathscr{P}'$, the random variable $\boldsymbol{Y}_N$ is understood as a vector of independent Poisson random variables with intensities $(np_1, \dots, np_S, 0, \dots)$, with $\sum_{j=1}^S p_j$ not necessarily equal to one, and $(P, n, \lambda) \mapsto \bar{\tau}_1(P, n, \lambda)$ is extended trivially from $\mathscr{P}$ to $\mathscr{P}'$ by letting $\bar{\tau}_1(P, n, \lambda) := n \sum_{j=1}^S p_j e^{-n(1+\lambda)p_j}$, $P \in \mathscr{P}'$. Then we have the following proposition, proved in Section S5.2.

PROPOSITION S2. *Let define* $n' \coloneqq (1+\delta)n$ *and let* S, ξ, δ *as defined previously. Then,* $\mathscr{E}(\lambda, n)$ *is bounded from below by*

$$\frac{1}{4n^2} \inf_{\hat{\rho}} \sup_{P \in \mathscr{P}'} \mathbb{E}_P^{n',\lambda}[(\bar{\tau}_1(P, n, \lambda) - \hat{\rho}(\boldsymbol{Y}_N))^2] - \frac{1}{n} - \Big(1 + \frac{n\xi(1+\lambda)}{S(1-\delta)}\Big)^2 \delta^2.$$

This implies that for any $\varepsilon > 0$, $\mathscr{E}(\lambda, n)$ *is bounded from below by*

$$\frac{\varepsilon^2}{4} \inf_{\hat{\rho}} \sup_{P \in \mathscr{P}'} \mathbb{P}_P^{n',\lambda}\big(|\bar{\tau}_1(P, n, \lambda) - \hat{\rho}(\boldsymbol{Y}_N)| > n\varepsilon\big) - \frac{1}{n} - \Big(1 + \frac{n\xi(1+\lambda)}{S(1-\delta)}\Big)^2 \delta^2.$$

Step 4. The next step involves applying the *method of the two fuzzy hypotheses* [14] to the result of Proposition S2. The next lemma is an adaptation of [14, Section 2.7.4] to our setting. Its proof is to be found in Section S5.3.

LEMMA S1 (Method of the two fuzzy hypotheses). *Let* $\mathcal{M}(\mathbb{N})$ *denote the space of all measures on* $\mathbb{N}$, *endowed with canonical* σ*-algebra. Let* $Q_1 = \sum_{j=1}^S q_{1,j}\delta_j$ *and* $Q_2 = \sum_{j=1}^S q_{2,j}\delta_j$ *be independent random variables taking values in* $\mathcal{M}(\mathbb{N})$. *Also let* $\mathscr{P}'$ *and* ε *as defined previously. Assume that for some* $0 < \alpha, \beta, \gamma < 1$ *with* $2\alpha + 2\beta + \gamma \leq 1$ *and with* n' *defined as above the following hold:*

1. $\mathbb{P}(Q_1 \notin \mathscr{P}') \leq \alpha$ *and* $\mathbb{P}(Q_2 \notin \mathscr{P}') \leq \alpha$;
2. $\mathbb{P}(|\bar{\tau}_1(Q_j, n, \lambda) - \mathbb{E}[\bar{\tau}_1(Q_j, n, \lambda)]| > n\varepsilon/2) \leq \beta$ *for* $j = 1, 2$;
3. $\mathbb{E}[\bar{\tau}_1(Q_1, n, \lambda)] \geq \mathbb{E}[\bar{\tau}_1(Q_2, n, \lambda)] + n\varepsilon$;
4. $\mathsf{TV}(\mathbb{E}[\otimes_{j=1}^S \mathrm{Poiss}(n' q_{1,j})], \mathbb{E}[\otimes_{j=1}^S \mathrm{Poiss}(n' q_{2,j})]) \leq \gamma$. *Here* $\mathsf{TV}(P, Q)$ *is used to denote the total-variation distance between probability measures* P *and* Q.

Then,

$$\inf_{\hat{\rho}} \sup_{P \in \mathscr{P}'} \mathbb{P}_P^{n',\lambda}(|\bar{\tau}_1(P, n, \lambda) - \hat{\rho}(\boldsymbol{Y}_N)| > n\varepsilon) \geq \frac{1}{2}\Big(1 - 2\alpha - 2\beta - \gamma\Big).$$

Step 5. The next step consists on constructing the hypotheses that will be used in conjunction with Lemma S1 and Proposition S2 to establish the minimax lower bound. The construction relies on ideas from [16, 15].

For some $L \in \mathbb{N}$ to be determined later, but satisfying $L \leq K_1 \xi$ for some constant $K_1 > 0$, we let U and V be two random variables taking values in $[0, \xi S^{-1}]$ such that $\mathbb{E}[U] = \mathbb{E}[V] = S^{-1}$ and when n is large enough,

$$\mathbb{E}[U^k] = \mathbb{E}[V^k] \quad \forall k \in \{0, \dots, L+1\},$$

$$\mathbb{E}[Ue^{-n(1+\lambda)U}] \geq \mathbb{E}[Ve^{-n(1+\lambda)V}] + S^{-1}\varepsilon.$$

The existence of such random variables is guaranteed by Lemma S2 below, proven in Section S5.4, for the appropriate choice of S, ξ, L and ε.

LEMMA S2. *Let $L \in \mathbb{N}$ and $\xi > 0$ such that $L \leq K_1\xi$ for some $K_1 > 0$. Let $S = \lceil n(1+\lambda) \rceil$. Then there exists $K_2 > 0$ (depending only on K_1) and two random variables U and V taking values in $[0, \xi S^{-1}]$ such that,*

$$\begin{gathered}\mathbb{E}[U^k] = \mathbb{E}[V^k] \quad \forall k \in \{0, \dots, L+1\},\\ \mathbb{E}[U] = \mathbb{E}[V] = S^{-1}, \quad \mathrm{Var}(U) \leq \xi S^{-2}, \quad \mathrm{Var}(V) \leq \xi S^{-2},\\ \mathbb{E}[Ue^{-n(1+\lambda)U}] \geq \mathbb{E}[Ve^{-n(1+\lambda)V}] + S^{-1}K_2 \min\left\{1, \sqrt{\xi/L^2}\exp(-L^2/\xi)\right\}.\end{gathered}$$

Then we let $(U_1, \dots, U_S)$, respectively $(V_1, \dots, V_S)$, be an independent vector of i.i.d. copies of U, respectively V, and we let

$$Q_1 = \sum_{j=1}^{S} U_j\delta_j, \quad \text{and,} \quad Q_2 = \sum_{j=1}^{S} V_j\delta_j.$$

The next proposition establishes conditions under which Q_1 and Q_2 as defined above meet the criteria of Lemma S1. The first two items are consequences of Bernstein's and Hoeffding's inequalities (respectively), item 3 is straightforward, and the last item is an immediate corollary of [16, Lemma 6]. The proof is given in Section S5.5.

PROPOSITION S3. *The following items are true.*

1. *Assume that $\mathrm{Var}(U) \leq \xi S^{-2}$, $\mathrm{Var}(V) \leq \xi S^{-2}$, and $S\delta^2 \geq 2\xi(1 + \delta/3)\log(2/\alpha)$. Then $\mathbb{P}(Q_1 \notin \mathscr{P}') \leq \alpha$ and $\mathbb{P}(Q_2 \notin \mathscr{P}') \leq \alpha$.*
2. *Assume that $S\varepsilon^2 \geq 2\xi\log(2/\beta)$. Then $\mathbb{P}(|\bar{\tau}_1(Q_1, n, \lambda) - \mathbb{E}[\bar{\tau}_1(Q_1, n, \lambda)]| > n\varepsilon/2) \leq \beta$. The same is also true for Q_2.*
3. *$\mathbb{E}[\bar{\tau}_1(Q_1, n, \lambda)] \geq \mathbb{E}[\bar{\tau}_1(Q_2, n, \lambda)] + n\varepsilon$.*
4. *Assume that $2\log(2)LS \geq n\xi(1+\delta)$ and $\gamma(2S)^{L+2}(L+2)! \geq 4S(n\xi(1+\delta))^{L+2}$. Then $\mathsf{TV}(\mathbb{E}[\otimes_{j=1}^{S}\mathrm{Poiss}(n'U_j)], \mathbb{E}[\otimes_{j=1}^{S}\mathrm{Poiss}(n'V_j)]) \leq \gamma$.*

Step 6. The proof of Theorem 3 follows from combining Propositions S2, S3, and Lemma S1, by choosing the constants α, β, γ and variables $\varepsilon, S, \xi, \delta, L$ accordingly. We now make explicit the choice for these constants and variables.

In the following for any $x > 0$ the notations $\lceil x \rceil$ stands for the smallest integer greater or equal than x. Then, for constants $c_0, c_1 > 0$ to be

determined we choose

$$S = \lceil n(1+\lambda) \rceil, \tag{S5}$$

$$\delta = c_0 \varepsilon / \xi, \tag{S6}$$

$$\xi = (2c_1/e) \min\{(1+\lambda)\log n,\, \log^2 n\}. \tag{S7}$$

For another constant $c_2 > 0$ to be determined, we further define $A(\lambda, n) > 0$ to be the solution to

$$A(\lambda,n)\log A(\lambda,n) = c_1^{-1} + c_1^{-1}\frac{\log(1+\lambda) - (1/2)\log\log(n) + \log(c_2)}{\log(n)}.$$

Then we pick (remark that this ensures that $L \leq K_1 \xi$ for some $K_1 > 0$, as requested previously),

$$L = \begin{cases} \lceil 2c_1 \log(n) \rceil & \text{if } 1+\lambda > \log(n), \\ \lceil c_1 A(\lambda,n) \log(n) \rceil & \text{if } 1+\lambda \leq \log(n), \end{cases} \tag{S8}$$

and for $c_3 > 0$ to be determined,

$$\varepsilon = c_3 \cdot \begin{cases} 1 & \text{if } 1+\lambda > \log(n), \\ \frac{1}{\sqrt{\log(n)}} \cdot \sqrt{\frac{2(1+\lambda)}{ec_1 A(\lambda,n)^2}} \cdot n^{-\frac{ec_1 A(\lambda,n)^2}{2(1+\lambda)}} & \text{if } 1+\lambda \leq \log(n). \end{cases} \tag{S9}$$

With this choice, we obtain the next proposition, proved in Section S5.6.

PROPOSITION S4. *Let $\alpha = \beta = \gamma = 1/10$, and let $S, \xi, \delta, L, \varepsilon$ as in Equations S5, S6, S7, S8 and S9. Then,*

1. $(1 + \frac{n\xi(1+\lambda)}{S(1-\delta)})^2 \delta^2 \leq c_0^2 \varepsilon^2 (1 + o(1))$ *as* $n \to \infty$;
2. *If* $\liminf_n \left\{ \frac{1+\lambda}{ec_1 A(\lambda,n)^2} \right\} > 1$ *then there exists* $n_0 > 0$ *such that for all* $n \geq n_0$ *it holds* $S\delta^2 \geq 2\xi(1+\delta/3)\log(2/\alpha)$;
3. *If* $\liminf_n \left\{ \frac{1+\lambda}{ec_1 A(\lambda,n)^2} \right\} > 1$ *then there exists* $n_0 > 0$ *such that for all* $n \geq n_0$ *it holds* $S\varepsilon^2 \geq 2\log(2/\beta)$;
4. *For any* $K_2 > 0$ *the constant* $c_3 > 0$ *can be chosen such that* $\varepsilon \leq K_2 \min\{1, \sqrt{\xi/L^2}\exp(-L^2/\xi)\}$; *In conjunction with Lemma S2 this guarantees the existence of* U *and* V *used in Step 5.*
5. *If* $c_2 > 0$ *is large enough, then there exists* $n_0 > 0$ *such that for all* $n \geq n_0$ *we have* $2\log(2)LS \geq n\xi(1+\delta)$ *and* $\gamma(2S)^{L+2}(L+2)! \geq 4S(n\xi(1+\delta))^{L+2}$.

Therefore, as a consequence of Propositions S2, S3 and Lemma S1, when c_0, c_1, c_2, c_3 are appropriately chosen, if $1+\lambda > ec_1 A(\lambda,n)^2$, and if n gets large enough,

$$\mathscr{E}(\lambda,n) \geq \Big(\frac{1}{16} - c_0^2 + o(1)\Big)\varepsilon^2.$$

Step 7. In view of Equation S9, the choice of c_1 shall be made cautiously. Indeed, the next proposition shows that $c_1 = 1/e$ is the optimal choice. The result of the next proposition also allows to get the final expression for the lower bound in $\mathscr{E}(\lambda, n)$, thus finishing the proof of Theorem 3. The proof of Proposition S5 is to be found in Section S5.7.

PROPOSITION S5. *Let $c_1 = 1/e$. Then whenever $1 + \lambda \leq \log(n)$ we have $A(\lambda, n) = e + o(1)$ as $n \to \infty$. Furthermore when $1 + \lambda \leq \log(n)$, as $n \to \infty$,*

$$c_1 A(\lambda, n)^2 \log(n) \leq e \log(n) + e \log \frac{c_2(1+\lambda)}{\sqrt{\log(n)}} + o(1).$$

S3. Numerical illustrations. We present an illustration on synthetic data of the estimators introduced in Section 2. We also consider other estimators of τ_1 that have been proposed in the literature of disclosure risk assessment: i) two parametric empirical Bayes estimators of τ_1 proposed by Bethlehem et al. [1] and Skinner et al. [13]; ii) a naive nonparametric estimator of τ_1; iii) a Bayesian nonparametric estimator of τ_1 proposed by Samuels [12]. A common feature of these estimators, as well as our class of nonparametric estimators, is that they rely on the Poisson abundance model for modeling the random partition induced by the cross-classified sample records. More recent approaches, not considered here, focus on modeling associations among identifying variables by log-linear models, local smoothing polynomials and hierarchical latent models. E.g., Manrique-Vallier and Reiter [8], Manrique-Vallier and Reiter [9], Carota et al. [2] and Carota et al. [3]. In particular, the Bayesian hierarchical semiparametric models of Carota et al. [2] and Carota et al. [3] show a remarkable better performance than models for random partitions, at the cost of an increasing computational effort for the need of Markov chain Monte Carlo methods for posterior approximation.

The approach of Bethlehem et al. [1] is a parametric empirical Bayes approach in the sense of Efron and Morris [4]. It relies on the following modeling assumption for the cells' frequencies of the population: $Y_j(\boldsymbol{X}, \bar{n}) \sim \text{Poiss}(\bar{n} p_j)$, where $\bar{n}$ is the size of the entire population. Bethlehem et al. [1] also assumed a Gamma prior distribution over the probabilities associated to each cell, namely $p_j \sim \text{Gam}(\alpha, \beta)$. One should specify the p_j's under the condition $\sum_{j=1}^{K_{\bar{n}}} p_j = 1$, however, for the sake of simplicity, Bethlehem et al. [1] assumed that $\sum_{j=1}^{K_{\bar{n}}} \mathbb{E}[p_j] = 1$, which is tantamount to saying that $\alpha = 1/(K_{\bar{n}}\beta)$. Under these modeling assumptions, Bethlehem et al. [1] proposed an estimator of the expected value of total number $T_1(\boldsymbol{X}, \bar{n})$ of population

uniques, i.e.,

$$T_1(\boldsymbol{X}, \bar{n}) := \sum_{j=1}^{K_{\bar{n}}} \mathbb{1}_{\{Y_j(\boldsymbol{X},\bar{n})=1\}}. \tag{S10}$$

Under the above Poisson-Gamma model, $\mathbb{E}[T_1(\boldsymbol{X}, \bar{n})] = \bar{n}(1+\bar{n}\beta)^{-(1+\alpha)}$, which depends on the parameters α and β, with the condition $\alpha = 1/(K\beta)$. Parameters can be easily estimated via maximum likelihood, as we have done in the subsequent numerical experiments. If $K_{\bar{n}}$ is not available, Bethlehem et al. [1] suggested to estimate $K_{\bar{n}}$ assuming a uniform distribution over the cells, hence

$$\hat{K}_{\bar{n}} = \frac{\bar{n}K_n}{\sum_{j=1}^{K_n} \mathbb{1}_{\{Y_j(\boldsymbol{X},n)=1\}}},$$

where n is the size of the observed sample and K_n stands for the number of distinct cells dictated by the sample of size n. If $\hat{\alpha}$ and $\hat{\beta}$ denote the maximum likelihood estimators of α and β, respectively, then an estimator of $T_1(\boldsymbol{X}, \bar{n})$ is $\hat{T}_1 = \bar{n}(1+\bar{n}\hat{\beta})^{-(1+\hat{\alpha})}$. Bethlehem et al. [1] then suggested a corresponding estimator of τ_1 as the sample portion of $\hat{T}_1$. More precisely, they proposed

$$\hat{\tau}_1^B = \frac{n}{\bar{n}}\hat{T}_1 = n(1+\bar{n}\hat{\beta})^{-(1+\hat{\alpha})}. \tag{S11}$$

as an estimator of τ_1. Skinner et al. [13] improved the estimator (S11). In particular, still under the Poisson-Gamma model, they considered directly the problem of estimating τ_1. In particular, they proposed the following estimator

$$\hat{\tau}_1^S := K_n \left(\frac{1+\bar{n}\hat{\beta}}{1+n\hat{\beta}}\right)^{-(1+\hat{\alpha})}, \tag{S12}$$

where the prior parameters α and β can be estimated via maximum likelihood. The estimators proposed in Section 2, due to their nonparametric empirical Bayes interpretation in the sense of Robbins [11], may be considered as the natural nonparametric counterparts of the empirical Bayes estimator (S12).

Besides parametric estimators of τ_1, we also consider two nonparametric estimators. A naive nonparametric estimator of τ_1 relies on the intuition that a natural estimator of τ_1 is the sampling fraction, with respect to the population, of the number of sample uniques. This estimator was first discussed in Bethlehem et al. [1] and Skinner et al. [13], and it is defined as

follows

$$\hat{\tau}_1^{\mathscr{N}} := Z_1(\boldsymbol{X}, n)\frac{n}{\bar{n}}. \tag{S13}$$

Samuels [12] exploits Bayesian nonparametric ideas, and in particular a Dirichlet process prior (Ferguson [5]) on the p_j's to derive a smoothed version of the naive estimator (S13). In particular, Samuels [12] suggested the following estimator

$$\hat{\tau}_1^{\mathscr{D}} := Z_1(\boldsymbol{X}, n)\frac{n+\vartheta-1}{\bar{n}+\vartheta-1}, \tag{S14}$$

where ϑ is the concentration parameter of the Dirichlet process prior. It is well-known (see, e.g. Ferguson [5]) that the maximum likelihood estimator of ϑ can be obtained by solving, with respect to ϑ, the equation $K_n = \sum_{1\leq j\leq n-1}\vartheta/(\vartheta+j)$.

We study the behavior of the Normalized Mean Squared Error (NMSE), with respect to the sampling fraction $(1+\lambda)^{-1}$, for the collection of estimators of τ_1 introduced before. In order to do that, we generate a collection of synthetic tables with C cells, where $C = 3\cdot 10^6$ in all our experiments. The population size is fixed to $\bar{n} = 10^6$, and we evaluate the NMSE for different values of the sample size $n = \bar{n}(\lambda+1)^{-1}$. The true probabilities $(p_j)_{j\geq 1}$ of cells are generated according to different types of distributions: the Zipf distribution, i.e., $p_j \propto j^{-s}$ for some $s > 0$, the uniform distribution over the total number of cells and the uniform Dirichlet distribution. Each Figure corresponds to a different choice of the distribution over the cells' probabilities: the Zipf distribution with respective parameter $s = 0.6, 0.8, 1$ (Figures S1–S3), the uniform distribution (Figure S4), the uniform Dirichlet distribution with respective parameter $\beta = 0.5, 1$ (Figures S5–S6). Each figure shows how the NMSE varies as a function of the sampling fraction $(1+\lambda)^{-1}$ for different estimators: i) the nonparametric estimator with Binomial smoothing $\hat{\tau}_1^{L_b}$, see Proposition 2; ii) the nonparametric estimator with Poisson smoothing $\hat{\tau}_1^{L_p}$, see Proposition 1; iii) the naive nonparametric estimator $\hat{\tau}_1^{\mathscr{N}}$; iv) the Bayesian nonparametric estimator $\hat{\tau}_1^{\mathscr{D}}$; v) the parametric empirical Bayes estimator $\hat{\tau}_1^{B}$; vi) the parametric empirical Bayes estimator $\hat{\tau}_1^{S}$. All experiments are averaged over 100 iterations and the empirical bands represent one standard deviation from the mean of the corresponding estimates.

The sampling fractions considered in our simulation study are above the limiting threshold $(\log n)^{-1}$. Within this range of sampling fractions, we do not observe a clear behavior for the performance of the estimators. It is apparent that in most of the simulated scenarios our estimator outperforms

as the sampling fraction $(1+\lambda)^{-1}$ increases from the limiting threshold $(\log n)^{-1}$. From Figure S5, the Bayesian nonparametric estimator $\hat{\tau}_1^{\mathscr{D}}$ provides the smallest NMSE; this behaviour is not surprising since data are drawn from a Dirichlet distribution. In Figures S1–S3, better performances are achieved by the estimators $\hat{\tau}_1^{L_b}$ and $\hat{\tau}_1^{L_p}$. We further observe that the choice of the smoothing distribution L for $\hat{\tau}_1^{L}$, i.e. the Binomial smoothing or the Poisson smoothing, is crucial with respect to the performance of the corresponding estimators. In all the simulated scenarios the Binomial smoothing displays a better performance than the Poisson smoothing. Finally in Table S1, we report the estimates of τ_1 (with empirical confidence bands) when $(\lambda+1)^{-1} = 1/5$ for all the choices of the cells' probabilities, from the left to right: the Zipf distribution with parameter $s = 0.6, 0.8, 1$, the uniform distribution, the uniform Dirichlet distribution with parameter $\beta = 0.5, 1$. All experiments are averaged over 100 iterations and the empirical intervals represent one standard deviation from the mean of the corresponding estimates. From Table S1, we can deduce similar considerations as before.

S4. Tightness of the approximation lower bound. We show that a suitable Chebychev polynomial approximation of the exponential function achieves (up to a multiplicative constant) the lower bound of Lemma 2 in the main document.

In view of Section B.2 in the main document, letting $\gamma_C : [-1,1] \to \mathbb{R}$ such that $\gamma_C(x) \coloneqq e^{-C(x+1)}$, it is enough to find a sequence of polynomial $(q_L)_{L\geq 1}$ such that q_L has degree at most L and for a constant $K > 0$,

$$L \leq \sqrt{C} \Longrightarrow \sup_{x\in[-1,1]} |\gamma_C(x) - q_L(x)| \leq K, \tag{S15}$$

and,

$$\sqrt{C} \leq L \leq \zeta C \Longrightarrow \sup_{x\in[-1,1]} |\gamma_C(x) - q_L(x)| \leq K\frac{L}{\sqrt{C}} e^{-C\varphi(L/C)}, \tag{S16}$$

at least when C is large enough, and with φ defined in Equation (18) in the main document. If $L \leq \sqrt{C}$, then we pick $q_L(x) = 0$ identically, so that the equation (S15) is trivially satisfied with any $K \geq 1$, because $|\gamma_C(x)| \leq 1$. Thus it suffices to establish (S16). For any $k \geq 0$, we let $T_k : [-1,1] \to \mathbb{R}$ the k-th order Chebychev polynomial, defined uniquely through the equality $T_k(\cos(\theta)) = \cos(k\theta)$, for all $\theta \in [-\pi, \pi]$. Then, we choose,

$$q_L(x) \coloneqq \sum_{k=0}^{L} a_k(C)\cdot T_k(x), \quad a_k(C) \coloneqq \int_{-1}^{1} \frac{e^{-C(x+1)}T_k(x)}{\sqrt{1-x^2}}\mathrm{d}x. \tag{S17}$$

We collect in the next Lemma several facts about the polynomial q_L and its coefficients $a_k(C)$ which will be used to derive the rate of approximation of q_L to γ_C, in the uniform norm.

LEMMA S3. *The following items are true.*

1. $a_k(C) = \pi(-1)^k e^{-C} I_k(C)$ *for all* $k \geq 0$, *where* I_k *is the modified Bessel function of the first kind (see [10, pg. 248]).*
2. *The series* $q_\infty := \sum_{k=0}^{\infty} a_k(C) T_k$ *converges uniformly in* $[-1, 1]$, *and* $q_\infty(x) = \gamma_C(x)$ *for all* $x \in [-1, 1]$.
3. *For all* $D > 0$ *there exists* $B_0 > 0$ *such that for all* $B \geq B_0$ *and for all* $k \geq \max\{BC, 2\}$, *we have the bound* $|a_k(C)| \leq e^{-Dk}$.
4. *For all* $B > 0$ *there exists* $C_0 > 0$ *such that for all* $C > C_0$, *for all* $\sqrt{C} \leq L < k \leq BC$, *we have the bound*

$$|a_k(C)| \leq \sqrt{2\pi} \cdot \frac{\exp\{-C\varphi(k/C)\}}{\sqrt{C}}.$$

Using the results of the previous lemma, we obtain the following corollary on the error of the best uniform polynomial approximation to γ_C on $[-1, 1]$, written $E_L(\gamma_C, [-1, 1])$.

COROLLARY S1. *For all* $\zeta > 0$ *there exists* $C_0 > 0$ *such that for all* $C > C_0$ *and for all* $\sqrt{C} \leq L \leq \zeta C$

$$E_L(\gamma_C, [-1, 1]) \leq \sqrt{4\pi(1 + \zeta^2)} \cdot \frac{\sqrt{C}}{L} e^{-C\varphi(L/C)}.$$

Furthermore, the polynomial q_L *defined in* (S17) *achieves the previous upper bound; and in view of Lemma 2 and Section B.2 in the main document, this bound is the best possible, up to a multiplicative constant.*

S5. Remaining proofs for the minimax lower bound. This section gather all the proofs of the propositions and lemma stated in Section S2.

S5.1. *Proof of Proposition S1.* Using Jensen's inequality we deduce that

$$\begin{aligned}\mathscr{E}(\lambda, n) &= \inf_{\hat{\rho}} \sup_{P \in \mathscr{P}} n^{-2} \mathbb{E}_P^{n,\lambda}[\mathbb{E}_P^{n,\lambda}[(\tau_1(\boldsymbol{X}, N, M) - \hat{\rho}(\boldsymbol{Y}(\boldsymbol{X}, N)))^2 \mid \boldsymbol{Y}(\boldsymbol{X}, N)]] \\ &\geq \inf_{\hat{\rho}} \sup_{P \in \mathscr{P}} n^{-2} \mathbb{E}_P^{n,\lambda}[(\mathbb{E}_P^{n,\lambda}[\tau_1(\boldsymbol{X}, N, M) \mid \boldsymbol{Y}(\boldsymbol{X}, N)] - \hat{\rho}(\boldsymbol{Y}(\boldsymbol{X}, N)))^2].\end{aligned}$$

Note that there is no explicit dependency on $\boldsymbol{X}$ and M anymore in the last display, but only on the random variable $(\boldsymbol{X}, N) \mapsto \boldsymbol{Y}(\boldsymbol{X}, N)$ which,

under P, is distributed as an infinite vector of independent Poisson random variables with parameters $(np_1, np_2, \dots)$. Besides observe also that $N = \sum_{j\geq 1} Y_j(\boldsymbol{X}, N)$. Let define

$$\begin{aligned}\tilde{\tau}_1(\boldsymbol{Y}_N, P, n, \lambda) &:= \mathbb{E}_P^{n,\lambda}[\tau_1(\boldsymbol{X}, N, M) \mid \boldsymbol{Y}(\boldsymbol{X}, N)] \\ &= \sum_{j\geq 1} \mathbb{1}_{\{Y_j(\boldsymbol{X},N)=1\}} \mathbb{E}_P^{n,\lambda}[\mathbb{1}_{\{Y_j(\boldsymbol{X},N+M)-Y_j(\boldsymbol{X},N)=0\}} \mid \boldsymbol{Y}(\boldsymbol{X}, N)].\end{aligned}$$

Remark that $(Y_j(\boldsymbol{X}, N+M)-Y_j(\boldsymbol{X}, N) : j \in \mathbb{N})$ is independent of $\boldsymbol{Y}(\boldsymbol{X}, N)$ and is a collection of independent Poisson random variables with intensities $(\lambda np_j : j \in \mathbb{N})$. Henceforth, we get

$$\tilde{\tau}_1(\boldsymbol{Y}_N, P, n, \lambda) = \sum_{j\geq 1} e^{-\lambda np_j} \mathbb{1}_{\{Y_j(\boldsymbol{X},N)=1\}}, \tag{S18}$$

and besides, since we abusively let $\boldsymbol{Y}_N$ denote the random variable $(\boldsymbol{X}, N) \mapsto \boldsymbol{Y}(\boldsymbol{X}, N)$,

$$\mathscr{E}(\lambda, n) \geq \inf_{\hat{\rho}} \sup_{P\in\mathscr{P}} n^{-2} \mathbb{E}_P^{n,\lambda}[(\tilde{\tau}_1(\boldsymbol{Y}_N, P, n, \lambda) - \hat{\rho}(\boldsymbol{Y}_N))^2]. \tag{S19}$$

We now trade $\tilde{\tau}_1(\boldsymbol{Y}_N, P, n, \lambda)$ for its expectation whowh we define as $\bar{\tau}_1(P, n, \lambda) := \mathbb{E}_P^{n,\lambda}[\tau_1(\boldsymbol{X}, N, M)]$. Recall that under P the vector $\boldsymbol{Y}_N$ is distributed as independent Poisson with parameters $(np_1, np_2, \dots)$. Hence,

$$\bar{\tau}_1(P, n, \lambda) = \sum_{j\geq 1} e^{-\lambda np_j} \mathbb{E}_P^{n,\lambda}[\mathbb{1}_{\{Y_j(\boldsymbol{X},N)=1\}}] = n \sum_{j\geq 1} p_j e^{-(1+\lambda)np_j}.$$

Similarly, for any $P \in \mathscr{P}$,

$$\begin{aligned}&\mathbb{E}_P^{n,\lambda}[(\tilde{\tau}_1(\boldsymbol{Y}_N, P, n, \lambda) - \bar{\tau}_1(P, n, \lambda))^2] \\ &\qquad = \sum_{j\geq 1} np_j e^{-(1+2\lambda)np_j} \left\{1 - np_j e^{-np_j}\right\} \leq n.\end{aligned} \tag{S20}$$

Thus from (S19) and Young's inequality, we find that

$$\begin{aligned}\mathscr{E}(\lambda, n) \geq \frac{1}{2n^2} \inf_{\hat{\rho}} \sup_{P\in\mathscr{P}} &\mathbb{E}_P^{n,\lambda}[(\bar{\tau}_1(P, n, \lambda) - \hat{\rho}(\boldsymbol{Y}_N))^2] \\ &- \frac{1}{n^2} \mathbb{E}_P^{n,\lambda}[(\tilde{\tau}_1(\boldsymbol{Y}_N, P, n, \lambda) - \bar{\tau}_1(P, n, \lambda))^2].\end{aligned}$$

That is using (S20),

$$\mathscr{E}(\lambda, n) \geq \frac{1}{2} \inf_{\hat{\rho}} \sup_{P\in\mathscr{P}} n^{-2} \mathbb{E}_P^{n,\lambda}[(\bar{\tau}_1(P, n, \lambda) - \hat{\rho}(\boldsymbol{Y}_N))^2] - n^{-1}.$$

S5.2. *Proof of Proposition S2.* For any $P \in \mathscr{P}'$ we let $\tilde{P}(\cdot) := P(\cdot)/P(\mathbb{N})$, so that $\tilde{P} \in \mathscr{P}$ is a probability measure. We write $\tilde{p}_j := p_j/P(\mathbb{N})$, $j \in \{1, \dots, S\}$. Furthermore we let $m(P) := n \sum_{j=1}^{S} p_j$. Then since $\boldsymbol{Y}_N$ is a vector of independent Poisson random variables, is clear that for any $P \in \mathscr{P}'$

$$\mathbb{E}_{\tilde{P}}^{n,\lambda}[(\bar{\tau}_1(\tilde{P}, n, \lambda) - \hat{\rho}(\boldsymbol{Y}_N))^2] = \mathbb{E}_{P}^{m(P),\lambda}[(\bar{\tau}_1(\tilde{P}, n, \lambda) - \hat{\rho}(\boldsymbol{Y}_N))^2]. \tag{S21}$$

We now choose $\hat{\tau}$ to be an estimator satisfying for some $\zeta > 0$

$$\sup_{P \in \mathscr{P}'} \mathbb{E}_{P}^{m(P),\lambda}[(\bar{\tau}_1(\tilde{P}, n, \lambda) - \hat{\tau}(\boldsymbol{Y}_N))^2] \leq \inf_{\hat{\rho}} \sup_{P \in \mathscr{P}'} \mathbb{E}_{P}^{m(P),\lambda}[(\bar{\tau}_1(\tilde{P}, n, \lambda) - \hat{\rho}(\boldsymbol{Y}_N))^2] + \zeta.$$

This is always possible for any $\zeta > 0$. Furthermore remark that $m(P) \leq (1+\delta)n = n'$, so that $m(P)/n' \leq 1$ always when $P \in \mathscr{P}'$. Let $P \in \mathscr{P}'$ be fixed, and let $\boldsymbol{W} = (W_1, W_2, \dots)$ such that conditional on $\boldsymbol{Y}_N$, the random variables W_j are independent binomial random variables with parameters $(Y_j, m(P)/n')$. Then define $\tilde{\tau}(\boldsymbol{Y}_N) := \mathbb{E}[\hat{\tau}(\boldsymbol{W})] \mid \boldsymbol{Y}_N]$. By Jensen's inequality,

$$\begin{aligned}
\mathbb{E}_{P}^{n',\lambda}[(\bar{\tau}_1(\tilde{P}, n, \lambda) - \tilde{\tau}(\boldsymbol{Y}_N))^2] &= \mathbb{E}_{P}^{n',\lambda}[(\mathbb{E}[\bar{\tau}_1(\tilde{P}, n, \lambda) - \hat{\tau}(\boldsymbol{W}) \mid \boldsymbol{Y}_N])^2] \\
&\leq \mathbb{E}_{P}^{n',\lambda}[\mathbb{E}[(\bar{\tau}_1(\tilde{P}, n, \lambda) - \hat{\tau}(\boldsymbol{W}))^2 \mid \boldsymbol{Y}_N]] \\
&= \mathbb{E}_{P}^{m(P),\lambda}[(\bar{\tau}_1(\tilde{P}, n, \lambda) - \hat{\tau}(\boldsymbol{Y}_N))^2] \\
&\leq \inf_{\hat{\rho}} \sup_{P \in \mathscr{P}'} \mathbb{E}_{P}^{m(P),\lambda}[(\bar{\tau}_1(\tilde{P}, n, \lambda) - \hat{\rho}(\boldsymbol{Y}_N))^2] + \zeta.
\end{aligned}$$

Taking the supremum over $P \in \mathscr{P}'$ on the lhs of the last display, and using that the infinimum over $\hat{\rho}$ will be always smaller than the value at $\tilde{\tau}$, we find using (S21) that

$$\begin{aligned}
&\inf_{\hat{\rho}} \sup_{P \in \mathscr{P}} \mathbb{E}_{P}^{n,\lambda}[(\bar{\tau}_1(P, n, \lambda) - \hat{\rho}(\boldsymbol{Y}_N))^2] \\
&\quad = \inf_{\hat{\rho}} \sup_{P \in \mathscr{P}'} \mathbb{E}_{\tilde{P}}^{n,\lambda}[(\bar{\tau}_1(\tilde{P}, n, \lambda) - \hat{\rho}(\boldsymbol{Y}_N))^2] \\
&\quad = \inf_{\hat{\rho}} \sup_{P \in \mathscr{P}'} \mathbb{E}_{P}^{m(P),\lambda}[(\bar{\tau}_1(\tilde{P}, n, \lambda) - \hat{\rho}(\boldsymbol{Y}_N))^2] \\
&\quad \geq \inf_{\hat{\rho}} \sup_{P \in \mathscr{P}'} \mathbb{E}_{P}^{n',\lambda}[(\bar{\tau}_1(\tilde{P}, n, \lambda) - \hat{\rho}(\boldsymbol{Y}_N))^2] - \zeta.
\end{aligned}$$

Since the previous is true for all $\zeta > 0$, we indeed have proven

$$\begin{aligned}
&\inf_{\hat{\rho}} \sup_{P \in \mathscr{P}} \mathbb{E}_{P}^{n,\lambda}[(\bar{\tau}_1(P, n, \lambda) - \hat{\rho}(\boldsymbol{Y}_N))^2] \\
&\qquad \geq \inf_{\hat{\rho}} \sup_{P \in \mathscr{P}'} \mathbb{E}_{P}^{n',\lambda}[(\bar{\tau}_1(\tilde{P}, n, \lambda) - \hat{\rho}(\boldsymbol{Y}_N))^2].
\end{aligned} \tag{S22}$$

To finish the proof of the proposition, we will now show that $\bar{\tau}_1(\tilde{P}, n)$ in (S22) can be traded for $\bar{\tau}_1(P, n, \lambda)$ at small cost. Remark that by Young's inequality, for any $P \in \mathscr{P}'$ and any $\hat{\rho}$,

$$\begin{aligned}(\text{S23}) \quad & \mathbb{E}_P^{n',\lambda}[(\bar{\tau}_1(\tilde{P}, n, \lambda) - \hat{\rho}(\boldsymbol{Y}_N))^2] \\ & \geq \frac{1}{2}\mathbb{E}_P^{n',\lambda}[(\bar{\tau}_1(P, n, \lambda) - \hat{\rho}(\boldsymbol{Y}_N))^2] - (\bar{\tau}_1(P, n, \lambda) - \bar{\tau}_1(\tilde{P}, n, \lambda))^2,\end{aligned}$$

with

$$\begin{aligned}&\bar{\tau}_1(P, n, \lambda) - \bar{\tau}_1(\tilde{P}, n, \lambda) \\ &\quad = n\sum_{j=1}^{S}(\tilde{p}_j - p_j)e^{-(1+\lambda)np_j} - n\sum_{j=1}^{S}\tilde{p}_j e^{-(1+\lambda)np_j}\Big\{1 - e^{n(1+\lambda)(p_j - \tilde{p}_j)}\Big\}.\end{aligned}$$

Hence,

$$\begin{aligned}(\text{S24}) \quad & |\bar{\tau}_1(P, n, \lambda) - \bar{\tau}_1(\tilde{P}, n, \lambda)| \\ & \leq n\sum_{j=1}^{S}|\tilde{p}_j - p_j| + n\sum_{j=1}^{S}\tilde{p}_j e^{-(1+\lambda)np_j}|1 - e^{n(1+\lambda)(p_j - \tilde{p}_j)}|.\end{aligned}$$

The first term of the rhs of the last display is easily seen to be bounded by $n\delta$ since $|p_j - \tilde{p}_j| = \tilde{p}_j|\sum_{k=1}^{S} p_k - 1| \leq \delta\tilde{p}_j$ for all $j = 1, \ldots, S$. For the second term, we use that $0 \leq 1 - e^{-x} \leq x$ for all $x \geq 0$. Hence, if $p_j \leq \tilde{p}_j$ we have,

$$\begin{aligned}\tilde{p}_j e^{-(1+\lambda)np_j}|1 - e^{n(1+\lambda)(p_j - \tilde{p}_j)}| &= \tilde{p}_j e^{-(1+\lambda)np_j}(1 - e^{-n(1+\lambda)(\tilde{p}_j - p_j)}) \\ &\leq n(1+\lambda)|\tilde{p}_j - p_j| \cdot \tilde{p}_j \\ &\leq n\delta(1+\lambda) \cdot \tilde{p}_j^2,\end{aligned}$$

while if $p_j > \tilde{p}_j$

$$\begin{aligned}\tilde{p}_j e^{-(1+\lambda)np_j}|1 - e^{n(1+\lambda)(p_j - \tilde{p}_j)}| &= \tilde{p}_j e^{-(1+\lambda)n\tilde{p}_j}(1 - e^{-n(1+\lambda)(p_j - \tilde{p}_j)}) \\ &\leq n(1+\lambda)|\tilde{p}_j - p_j| \cdot \tilde{p}_j \\ &\leq n\delta(1+\lambda) \cdot \tilde{p}_j^2.\end{aligned}$$

Therefore in any cases the second term of the rhs of Equation (S24) is bounded above by $n^2\delta(1+\lambda)\sum_{j=1}^{S}\tilde{p}_j^2$, and thus

$$|\bar{\tau}_1(P, n, \lambda) - \bar{\tau}_1(\tilde{P}, n, \lambda)| \leq n\delta + n^2\delta(1+\lambda)\sum_{j=1}^{S}\tilde{p}_j^2 \leq \Big(1 + \frac{n\xi(1+\lambda)}{S(1-\delta)}\Big)n\delta.$$

This estimate combined with (S22) and (S23) completes the proof for the first inequality of the proposition. The second inequality simply follows from the first by an application of Markov's inequality.

S5.3. *Proof of Lemma S1.* The proof is a trivial adaptation of the classical Le Cam method with two fuzzy hypotheses, as also described in [14].

Let $\hat{\rho}$ be fixed but arbitrary and let define for convenience the events $A_n(P;\hat{\rho}) \coloneqq \{\boldsymbol{Y}_N : |\bar{\tau}_1(P,n,\lambda) - \hat{\rho}(\boldsymbol{Y}_N)| > n\varepsilon\}$. Since the average is always less or equal than the supremum over $\mathscr{P}'$, we establish that

$$\begin{aligned}
&\sup_{P\in\mathscr{P}'} \mathbb{P}_P^{n',\lambda}(A_n(P;\hat{\rho})) \\
&\quad\geq \frac{1}{2}\mathbb{E}\big[\mathbb{P}_{Q_1}^{n',\lambda}(A_n(Q_1;\hat{\rho}))\mathbf{1}_{\mathscr{P}'}(Q_1)\big] + \frac{1}{2}\mathbb{E}\big[\mathbb{P}_{Q_2}^{n',\lambda}(A_n(Q_2;\hat{\rho}))\mathbf{1}_{\mathscr{P}'}(Q_2)\big] \\
&\quad\geq \frac{1}{2}\mathbb{E}\big[\mathbb{P}_{Q_1}^{n',\lambda}(A_n(Q_1;\hat{\rho}))\big] + \frac{1}{2}\mathbb{E}\big[\mathbb{P}_{Q_2}^{n',\lambda}(A_n(Q_2;\hat{\rho}))\big] - \alpha,
\end{aligned}$$

where for the last line we have used the item 1 of the Lemma.

Now let define the events $B_n(Q_j;\hat{\rho}) \coloneqq \{\boldsymbol{Y}_N : |\mathbb{E}[\bar{\tau}_1(Q_j,n,\lambda)] - \hat{\rho}(\boldsymbol{Y}_N)| > n\varepsilon/2\}$, for $j = 1,2$. Under item 2 of the Lemma, it is rapidly obtained from the last display that

$$\begin{aligned}
&\sup_{P\in\mathscr{P}'} \mathbb{P}_P^{n',\lambda}(A_n(P;\hat{\rho})) \\
&\quad\geq \frac{1}{2}\mathbb{E}\big[\mathbb{P}_{Q_1}^{n',\lambda}(B_n(Q_1;\hat{\rho}))\big] + \frac{1}{2}\mathbb{E}\big[\mathbb{P}_{Q_2}^{n',\lambda}(B_n(Q_2;\hat{\rho}))\big] - \alpha - \beta \\
&\quad= \frac{1}{2}\mathbb{E}\big[1 - \mathbb{P}_{Q_1}^{n',\lambda}(B_n(Q_1;\hat{\rho})^c) + \mathbb{P}_{Q_2}^{n',\lambda}(B_n(Q_2;\hat{\rho}))\big] - \alpha - \beta.
\end{aligned}$$

But under item 3 of the lemma, we have that $B_n(Q_1;\hat{\rho})^c \subseteq B_n(Q_2;\hat{\rho})$. Moreover under Q_j, $j = 1,2$, $\boldsymbol{Y}_N$ is a vector of independent Poisson random variables with parameters $(n'q_{j,1},\dots,n'q_{j,S},0,\dots)$ and thus by the classical Le Cam's trick the last equation is bounded by

$$\begin{aligned}
&\sup_{P\in\mathscr{P}'} \mathbb{P}_P^{n',\lambda}(A_n(P;\hat{\rho})) \\
&\quad\geq \frac{1}{2}\Big(1 - \mathsf{TV}\big(\mathbb{E}[\otimes_{j=1}^{S}\mathrm{Poiss}(n'q_{1,j})],\, \mathbb{E}[\otimes_{j=1}^{S}\mathrm{Poiss}(n'q_{2,j})]\big)\Big) - \alpha - \beta \\
&\quad\geq \frac{1}{2}\Big(1 - \gamma - 2\alpha - 2\beta\Big),
\end{aligned}$$

where the last line follows from the item 4 of the Lemma. Since the rhs of the last display is independent of $\hat{\rho}$, the conclusion of the Lemma follows.

S5.4. *Proof of Lemma S2.* The proof of Lemma S2 follows the guidelines used in the papers Wu and Yang [16, 15], relating the problem of the existence of the random variables to the problem of finding the best polynomial approximation to some function.

For $a, b \in \mathbb{R}$, we let $\mathsf{C}[a,b]$ denote the space of continuous functions on $[a,b]$, and for any $L \in \mathbb{Z}_+$ we let $\mathsf{P}_L[a,b] \subset \mathsf{C}[a,b]$ denote the space of polynomials of degree no more than L on $[a,b]$. For any $f \in \mathsf{C}[a,b]$, the best polynomial (of degree at most L) approximation to f is defined as

$$E_L(f,[a,b]) := \inf\{\sup\{|f(x)-q(x)| : \ x \in [a,b]\} : \ q \in \mathsf{P}_L[a,b]\}.$$

For the sake of simplicity, we define $B := n(1+\lambda)\xi/(2S)$. We also define $g : [\xi^{-1}, 1] \to \mathbb{R}_+$ such that $g(x) := \exp\{-2Bx\}$. It is a classical result that for any $L \in \mathbb{N}$ we can find random variables X and Y taking values in $[\xi^{-1}, 1]$ and such that

$$\begin{aligned} \mathbb{E}[X^k] &= \mathbb{E}[Y^k], \qquad k = 0, \dots, L, \\ \mathbb{E}[g(X)] &= \mathbb{E}[g(Y)] + E_L(g, [\xi^{-1}, 1]). \end{aligned}$$

The proof of the existence of such random variables can be found for instance in Wu and Yang [15, 16] for a constructive argument, or for instance in Lepski et al. [6] using the Hahn-Banach theorem and a duality argument.

We now assume that we have at our disposal the random variables X and Y of the previous paragraph, and we write P_X and P_Y their distributions. The construction of the random variables U and V is done using the trick introduced in Wu and Yang [15, Lemma 4]. Namely, we let U and V having respective distributions on $[0, \xi S^{-1}]$

$$\begin{aligned} P_U(\mathrm{d}x) &:= \big(1 - \mathbb{E}[(\xi X)^{-1}]\big)\delta_0 + (Sx)^{-1} P_{\xi X/S}(\mathrm{d}x), \\ P_V(\mathrm{d}x) &:= \big(1 - \mathbb{E}[(\xi Y)^{-1}]\big)\delta_0 + (Sx)^{-1} P_{\xi Y/S}(\mathrm{d}x). \end{aligned}$$

Because $X, Y \geq \xi^{-1}$ almost-surely, then $\mathbb{E}[(\xi X)^{-1}] \leq 1$ and $\mathbb{E}[(\xi Y)^{-1})] \leq 1$. Indeed from Wu and Yang [15, Lemma 4], P_U and P_V are proper probability distributions on $[0, \xi S^{-1}]$ satisfying

$$\begin{aligned} &\mathbb{E}[U] = \mathbb{E}[V] = 1/S, \qquad \mathbb{E}[U^k] = \mathbb{E}[V^k], \qquad k = 0, \dots, L+1, \\ &\mathbb{E}[U \exp\{-n(1+\lambda)U\}] = \mathbb{E}[V \exp\{-n(1+\lambda)V\}] + S^{-1} E_L(g, [\xi^{-1}, 1]). \end{aligned}$$

Furthermore, it is clear that,

$$\mathbb{E}[U^2] = \frac{1}{S} \int x \, P_{\xi x/S}(\mathrm{d}x) = \frac{\xi \mathbb{E}[X]}{S^2} \leq \frac{\xi}{S^2}.$$

Hence $\mathrm{Var}(U) \leq \xi/S^2$. It is obvious that we also have $\mathrm{Var}(V) \leq \xi/S^2$. Thus, the proof of the theorem is finished by obtaining a lower bound on the best polynomial approximation $E_L(g, [\xi^{-1}, 1])$. This is a consequence of

the Lemma 2 in the main paper since $L \leq K_1\xi$, $B = (\xi/2)(1 + O(\xi^{-1}))$, and also because

$$\frac{\xi}{2}\varphi\Big(\frac{2L}{\xi}\Big) \leq \frac{\xi}{2} \cdot \frac{1}{2}\Big(\frac{2L}{\xi}\Big)^2 = \frac{L^2}{\xi},$$

by using the facts about φ derived in Section S6.3.

S5.5. *Proof of Proposition S3.* Here we prove separately all the items stated in Proposition S3.

Proof of item 1. The proof is an immediate consequence of Bernstein's inequality using that $\mathrm{Var}(U) \leq \xi S^{-2}$ and $0 \leq U \leq \xi S^{-1}$. Similarly for V. □

Proof of item 2. The proof for Q_1 and Q_2 are identical, thus we only prove the result for Q_1. By definition, we have that

$$\bar{\tau}_1(Q_1, n, \lambda) = n \sum_{j=1}^{S} U_j e^{-n(1+\lambda)U_j}.$$

Whence, $\bar{\tau}_1(Q_1, n, \lambda)$ is a sum of i.i.d. random variables taking values in $[0, n\xi S^{-1}]$. By Hoeffding's inequality,

$$\mathbb{P}\big(|\bar{\tau}_1(Q_1, n, \lambda) - \mathbb{E}[\bar{\tau}_1(Q_1, n, \lambda)]| > n\varepsilon/2\big) \leq 2\exp\Big\{-\frac{S\varepsilon^2}{2\xi}\Big\}.$$

The conclusion follows from simple algebraic manipulations. □

Proof of item 3. This is immediate by remarking that $\mathbb{E}[\bar{\tau}_1(Q_1, n, \lambda)] = nS\mathbb{E}[Ue^{-n(1+\lambda)U}]$ and $\mathbb{E}[\bar{\tau}_1(Q_2, n, \lambda)] = nS\mathbb{E}[Ve^{-n(1+\lambda)V}]$. □

Proof of item 4. Since $(U_1, \ldots, U_S)$ and $(V_1, \ldots, V_S)$ are independent and i.i.d vectors, we obtain immediately that

$$\begin{aligned}\text{(S25)}\quad &\mathsf{TV}(\mathbb{E}[\otimes_{j=1}^{S}\mathrm{Poiss}(n'U_j)], \mathbb{E}[\otimes_{j=1}^{S}\mathrm{Poiss}(n'V_j)]) \\ &\qquad = S\mathsf{TV}(\mathbb{E}[\mathrm{Poiss}(n'U)], \mathbb{E}[\mathrm{Poiss}(n'V)]).\end{aligned}$$

Since $0 \leq U, V \leq \xi S^{-1}$ almost-surely, we obtain from [16, Lemma 6],

$$\begin{aligned}&\mathsf{TV}\Big(\mathbb{E}[\mathrm{Poiss}(n'U)], \mathbb{E}[\mathrm{Poiss}(n'V)]\Big) \\ &\qquad \leq \frac{1}{(L+2)!}\Big(\frac{n'\xi}{2S}\Big)^{L+2}\Big(2 + 2^{n'\xi/(2S)-L} + 2^{n'\xi/(2\log(2)S)-L}\Big).\end{aligned}$$

Recall that $n' = n(1+\delta)$, thus under the conditions of the proposition we have $n'\xi/(2S) \le n'\xi/(2\log(2)S) \le L$, and hence from the last display we obtain that

$$\mathsf{TV}\Big(\mathbb{E}[\mathrm{Poiss}(n'U)], \mathbb{E}[\mathrm{Poiss}(n'V)]\Big) \le \frac{4}{(L+2)!}\Big(\frac{n\xi(1+\delta)}{2S}\Big)^{L+2}. \tag{S26}$$

Then the conclusion follows by combining Equations (S25) and (S26). □

S5.6. *Proof of Proposition S4.* Here we prove separately all the items stated in Proposition S4.

Proof of item 1. From the definitions of ξ, S and δ, we immediately see that $(1+\frac{n\xi(1+\lambda)}{S(1-\delta)})^2\delta^2 \le (1+\frac{\xi}{1-\delta})^2 c_0^2\varepsilon^2/\xi^2 = c_0^2\varepsilon^2(1+o(1))$, because $\xi \to \infty$ and $\varepsilon = O(1)$ (the latter fact is easier to see a posteriori). □

Proof of items 2 and 3. The case $1+\lambda > \log(n)$ is straightforward, thus we focus only on $1+\lambda \le \log(n)$. For the sake of simplicity, we define $r := \sqrt{\log(n)/(1+\lambda)}$ and $y := \sqrt{ec_1}A(\lambda,n)$, so that $\varepsilon = \sqrt{2}c_3(ry)^{-1}\exp(-r^2y^2/2)$. Then from the definitions of S, δ and ξ,

$$\begin{aligned} S\delta^2 &\ge c_0^2 n(1+\lambda)\frac{\varepsilon^2}{\xi^3}\cdot\xi \\ &\gtrsim n\cdot\max\Big\{\frac{1}{(1+\lambda)^2\log^3(n)}, \frac{1+\lambda}{\log^6(n)}\Big\}\varepsilon^2\cdot\xi \\ &\gtrsim \xi\cdot n\cdot\max\Big\{\frac{1}{(1+\lambda)^2\log^3(n)}, \frac{1+\lambda}{\log^6(n)}\Big\}\frac{1}{(ry)^2}e^{-r^2y^2}. \end{aligned}$$

But under the assumption of the Proposition, have $\liminf_n\big\{\frac{\log(n)}{r^2y^2}\big\} > 1$, which entails that for n large enough $S\delta^2 \ge 2\xi(1+\delta/3)\log(20)$. The proof of item 3 is similar. □

Proof of item 4. This is an immediate consequence of the definitions of ε, L and ξ. □

Proof of item 5, Case $1+\lambda \le \log(n)$. Note that in this case we have $\xi = (2c_1/e)(1+\lambda)\log(n)$ and $L = \lceil c_1A(\lambda,n)\log(n)\rceil$. For n large enough such that $0 < \delta \le e\log(2) - 1$ (this always happens, see for instance the remark in the proof of item 1), we have

$$\begin{aligned} 2\log(2)LS &\ge 2\log(2)c_1A(\lambda,n)\log(n)\cdot n(1+\lambda) \\ &= n\xi\cdot e\log(2)A(\lambda,n) \end{aligned}$$

$$\begin{aligned}&\geq n\xi \cdot e\log(2)\\&\geq n\xi(1+\delta),\end{aligned}$$

where the third line follows because $\lambda \geq 0$ and from the definition of $A(\lambda, n)$ by remarking that $a \log a \geq 0 \Rightarrow a \geq 1$.

Further, using that $(L+2)! \geq L^2 L!$, and because $L \leq K_1\xi$ implies that $(1+\delta)^{L+2} \lesssim (1+\delta)^L \leq e^{L\delta} \leq e^{K_1 c_1 \varepsilon} \lesssim 1$, we have

$$\begin{aligned}\frac{4S}{(L+2)!}\Big(\frac{n\xi(1+\delta)}{2S}\Big)^{L+2} &\lesssim \frac{S}{L^2}\Big(\frac{n\xi}{2S}\Big)^2 \frac{1}{L!}\Big(\frac{n\xi}{2S}\Big)^L\\&= \frac{S}{L^2}\Big(\frac{c_1\log(n)}{e}\Big)^2 \frac{1}{L!}\Big(\frac{c_1\log(n)}{e}\Big)^L\\&\lesssim \frac{S\log^2(n)}{L^{5/2}}\Big(\frac{c_1\log(n)}{L}\Big)^L,\end{aligned}$$

where the last line follows from Stirling's formula. Using the definitions of $S \lesssim n(1+\lambda)$, L and $A(\lambda, n)$, we deduce that

$$\begin{aligned}\frac{4S}{(L+2)!}\Big(\frac{n\xi(1+\delta)}{2S}\Big)^{L+2} &\lesssim \frac{n(1+\lambda)A(\lambda,n)^{-c_1 A(\lambda,n)\log(n)}}{A(\lambda,n)^{5/2}\log^{1/2}(n)}\\&= \frac{1}{c_2 A(\lambda,n)^{5/2}} \leq \frac{1}{c_2}.\end{aligned}$$

Therefore by choosing $c_2 > 0$ large enough we obtain that $\gamma(2S)^{L+2}(L+2)! \geq 4S(n\xi(1+\delta))^{L+2}$. □

Proof of item 5, Case $1+\lambda > \log(n)$. Note that in this case we have $\xi = (2c_1/e)\log^2(n)$ and $L = \lceil 2c_1\log(n)\rceil$. For n large enough such that $0 < \delta \leq 2e\log(2) - 1$ (this always happens, see for instance the remark in the proof of item 1), we have

$$\begin{aligned}2\log(2)LS &\geq 4c_1\log(2)n(1+\lambda)\log(n)\\&\geq 4c_1\log(n)n\log^2(n)\\&= n\xi \cdot 2e\log(2)\\&\geq n\xi(1+\delta).\end{aligned}$$

Proceeding along similar lines as for the case $1+\lambda \leq \log(n)$, it is easily found that as $n \to \infty$ we have

$$\frac{4S}{(L+2)!}\Big(\frac{n\xi(1+\delta)}{2S}\Big)^{L+2} \to 0,$$

and hence certainly that $\gamma(2S)^{L+2}(L+2)! \geq 4S(n\xi(1+\delta))^{L+2}$ when n gets large enough. □

S5.7. *Proof of Proposition S5.* We define the function $\varphi : \mathbb{R}_+ \to \mathbb{R}$ such that $\varphi(x) = x\log(x)$. When $1+\lambda \leq \log(n)$, it is clear that $A(\lambda, n)$ converges to the solution of $\varphi(x) = c_1^{-1} = e$, hence $A(\lambda, n) \to e$, which proves the first claim.

For the second claim, let define,

$$\Delta_n := e\frac{\log(1+\lambda) - (1/2)\log\log(n) + \log(c_2)}{\log(n)}.$$

For n large enough such that $\Delta_n > -1$, it is clear than $A(\lambda, n) \geq 0$. Furthermore, by a Taylor expansion of φ near $x = e$, we find that there is a $\bar{x}$ in the line segment between $A(\lambda, n)$ and e such that

$$\begin{aligned}\varphi(A(\lambda,n)) &= \varphi(e) + \varphi'(e)(A(\lambda,n) - e) + \frac{\varphi''(\bar{x})}{2}(A(\lambda,n) - e)^2 \\ &\geq \varphi(e) + \varphi'(e)(A(\lambda,n) - e),\end{aligned}$$

because $\varphi''(x) = 1/x > 0$ whenever $x > 0$. Since $\varphi(A(\lambda, n)) - \varphi(e) = \Delta_n$, $\varphi(e) = e$, and $\varphi'(e) = 2$, we deduce that for those n large,

$$0 \leq A(\lambda, n) \leq e + \Delta_n/2.$$

Therefore,

$$\begin{aligned}e^{-1}A(\lambda,n)^2\log(n) &\leq e\log(n) + \Delta_n\log(n) + \frac{\Delta_n^2\log(n)}{4e} \\ &= e\log(n) + e\log\frac{c_2(1+\lambda)}{\sqrt{\log(n)}} + o(1).\end{aligned}$$

This concludes the proof.

S6. Proofs related to the upper-bound on the best polynomial approximation. In this section, we give the proofs of the results stated in Section S4, regarding the construction of a polynomial of degree no more than L achieving the approximation error of the Lemma 2 in the main document.

S6.1. *Proof of Lemma S3.* Below we prove the items stated in Lemma S3. The proofs mainly consist on driving the formula for $a_k(C)$ and getting sharp estimates on $|a_k(C)|$ for various regimes governed by the ratio k/C.

Proof of item (1). By doing the change of variable $x \mapsto \cos(\theta)$ in the definition of $a_k(C)$, and using that $T_k(\cos\theta) = \cos(k\theta)$ we obtain

$$a_k(C) = e^{-C}\int_{-1}^{1}\frac{e^{-Cx}T_k(x)}{\sqrt{1-x^2}}\mathrm{d}x$$

$$= e^{-C} \int_0^\pi e^{-C\cos\theta} \cos(k\theta) \mathrm{d}\theta$$
$$= \pi(-1)^k e^{-C} I_k(C),$$

where we used [10, formula 10.32.3]. □

PROOF OF ITEM (2). The uniform convergence of the series is an immediate consequence of the fact that $|T_n(x)| \leq 1$ for all $x \in [-1, 1]$ and the upper bound estimate on $|a_k(C)|$ obtained just after in item (3). □

PROOF OF ITEM (3). To prove the item, we use the classical bound on the modified Bessel function obtained by Luke [7]. Indeed, for any $k \geq BC$, we have

$$\begin{aligned} 0 \leq \pi e^{-C} I_k(C) &\leq \frac{\pi}{k!}\Big(\frac{C}{2}\Big)^k \\ &\leq \frac{\pi}{\sqrt{2\pi k}}\Big(\frac{eC}{2k}\Big)^k \\ &\leq \sqrt{\frac{\pi}{2k}} \exp\Big\{ - k \log\Big(\frac{2B}{e}\Big)\Big\}, \end{aligned}$$

where the first line comes from Luke [7], and the second line by Stirling's approximation. For $k \geq 2$ we have $\pi/(2k) \leq 1$. Thus, it is enough to take $B_0 = e^{(1+D)}/2$, which concludes the proof. □

PROOF OF ITEM (4). We follow a similar path as in the Section B.3 of the main document. Indeed, we can remark by Stirling's formula that for any $p, k \geq 0$ we have

$$(p+k)! \geq \sqrt{2\pi}(p+k)^{p+k+1/2} e^{-(p+k)}, \qquad p! \geq \sqrt{2\pi} p^{p+1/2} e^{-p}.$$

Then, by defining $\phi_{z,k}(x)$ as in Section B.3 of the main document, we obtain the upper bound,

$$\begin{aligned} e^{-C} I_k(C) &\leq \frac{e^{-C}}{k!}\Big(\frac{C}{2}\Big)^k + \sum_{p\geq 1} \frac{1}{p!(p+k)!}\Big(\frac{C}{2}\Big)^{2p+k} \\ &\leq \frac{e^{-C}}{k!}\Big(\frac{C}{2}\Big)^k + \frac{1}{2\pi}\sum_{p\geq 1} \frac{\exp\{\phi_{C,k}(p)\}}{\sqrt{p(p+k)}}. \end{aligned} \tag{S27}$$

We consider the first term of the rhs of the previous display. By Stirling's formula, we have

$$\frac{e^{-C}}{k!}\Big(\frac{C}{2}\Big)^k \leq \frac{e^{-C}}{\sqrt{2\pi k}}\Big(\frac{eC}{2k}\Big)^k$$

$$= \frac{1}{\sqrt{2\pi k}} \exp\Big\{ -C + k \log\Big(\frac{eC}{2k}\Big)\Big\}$$
$$= \frac{1}{\sqrt{2\pi k}} \exp\{\phi_{C,k}(0)\},$$

where $\phi_{C,k}(0)$ is defined by extending $\phi_{C,k}$ at zero by continuity. We remark that,

$$\begin{aligned}
&\phi_{C,k}(0) - \phi_{C,k}(x_0) \\
&\quad = -2x_0 + x_0 \log x_0 - k \log k + (x_0 + k)\log(x_0 + k) - 2x_0 \log \frac{z}{2} \\
&\quad = -2x_0 - k \log k + k \log(x_0 + k) + x_0\Big(\log x_0 + \log(x_0 + k) - 2 \log \frac{z}{2}\Big) \\
&\quad = -2x_0 + k \log\Big(1 + \frac{x_0}{k}\Big) - x_0 \phi'_{C,k}(x_0) \\
&\quad = -2x_0 + k \log\Big(1 + \frac{x_0}{k}\Big) \\
&\quad \leq -x_0.
\end{aligned}$$

It follows,

$$\frac{e^{-C}}{k!}\Big(\frac{C}{2}\Big)^k \leq \frac{\exp\{\phi_{C,k}(x_0)\}}{\sqrt{C}} \cdot \sqrt{\frac{C}{2\pi k}} e^{-x_0} = \frac{\exp\{\phi_{C,k}(x_0)\}}{\sqrt{C}} \cdot o(1)$$

as $C \to \infty$, by remarking that $C/k \lesssim \sqrt{C}$ and that $k \leq BC$, hence $C/k \geq B^{-1}$ and $x_0 \geq B'k > B'\sqrt{C} \to \infty$ for a universal constant $B' > 0$.

We now consider the second term in the rhs of (S27). We let p_0 be the integer defined in Section B.3 of the main document, that is $x_0 < p_0 \leq x_0 + 1$ is integer and $\phi'_{z,k}(x_0) = 0$. Recall that $x_0 \geq B'k > B'\sqrt{C} \to \infty$ for a universal constant $B' > 0$. Let $G_1 > 0$ be a constant to be chosen accordingly later, and let $A_1 \in \mathbb{N}$ be the only integer such that

$$x_0 - G_1\sqrt{x_0 \log(x_0)} - 1 < A_1 \leq x_0 - G_1\sqrt{x_0 \log(x_0)}.$$

By the previous discussion, we have $1 < A_1 < x_0$ at least for L large enough. Similarly, we let $G_2 > 0$ a constant to be chosen accordingly, and we let $A_2 \in \mathbb{N}$ be the only integer such that

$$x_0 + G_2(1 + \sqrt{x_0}) \log(x_0) \leq A_2 < x_0 + G_2(1 + \sqrt{x_0}) \log(x_0) + 1.$$

Obviously $A_2 > x_0$. Then we decompose the sum in the rhs of (S27) as

$$\underbrace{\sum_{p=1}^{A_1} \frac{\exp\{\phi_{C,k}(p)\}}{\sqrt{p(p+k)}}}_{S_1} + \underbrace{\sum_{p=A_1+1}^{A_2} \frac{\exp\{\phi_{C,k}(p)\}}{\sqrt{p(p+k)}}}_{S_2} + \underbrace{\sum_{p>A_2} \frac{\exp\{\phi_{C,k}(p)\}}{\sqrt{p(p+k)}}}_{S_3}.$$

The conclusion of the proof follows by gathering the bounds for S_1, S_2, and S_3, which are derived in the paragraphs below, and by using that $\phi_{C,k}(x_0) = -C\varphi(k/C)$.

Bound on S_1. Let $p \in [1, A_1]$. We remark by a Taylor expansion that $\phi_{C,k}(p) = \phi_{C,k}(x_0) + \frac{1}{2}\phi''_{C,k}(\bar{p})(p - x_0)^2$ for some $\bar{p} \in (1, x_0)$. As for Section B.3 of the main document, we see that $\phi''_{C,k}(\bar{p}) \leq -1/x_0$. Therefore, remarking that $(p - x_0)^2 \geq G_1^2 x_0 \log(x_0)$ for any $1 \leq p \leq A_1$ (at least for L large enough),

$$\begin{aligned} S_1 &\leq \frac{\exp\{\phi_{C,k}(x_0)\}}{\sqrt{k}} \sum_{p=1}^{A_1} \exp\Big(- \frac{(p - x_0)^2}{2x_0} \Big) \\ &\leq \frac{\exp\{\phi_{C,k}(x_0)\}}{\sqrt{C}} \cdot A_1 \sqrt{\frac{C}{k}} x_0^{-G_1^2/2} \\ &= \frac{\exp\{\phi_{C,k}(x_0)\}}{\sqrt{C}} \cdot o(1), \end{aligned}$$

where the last line follows by choosing G_1 large enough, because $A_1 \leq x_0$, $C/k \lesssim \sqrt{C}$, and $x_0 \geq B'\sqrt{C} \to \infty$.

Bound on S_2. Let $A_1 < p \leq A_2$. Then, $|p - x_0| = O\big(\sqrt{x_0}\log(x_0)\big)$ as $C \to \infty$. Further, it is easily seen that, as $C \to \infty$,

$$\begin{aligned} \sup_{x \in [A_1, A_2]} |\phi'''_{C,k}(x)| &= \sup_{x \in [A_1, A_2]} \Big(\frac{1}{x^2} + \frac{1}{(x + k)^2} \Big) \\ &\leq 2 \sup_{x \in [A_1, A_2]} \frac{1}{x^2} \\ &= \frac{2(1 + o(1))}{x_0^2}. \end{aligned}$$

Therefore, by Taylor expansion, and as $C \to \infty$,

$$\begin{aligned} \phi_{C,k}(p) &= \phi_{C,k}(x_0) + \frac{1}{2}\phi''(x_0)(p - x_0)^2 + O\Big(\frac{x_0^{3/2} \log^3(x_0)}{x_0^2} \Big) \\ &= \phi_{C,k}(x_0) + \frac{1}{2}\phi''(x_0)(p - x_0)^2 + o(1). \end{aligned}$$

It follows,

$$S_2 \leq (1 + o(1)) \cdot \frac{\exp\{\phi_{C,k}(x_0)\}}{\sqrt{A_1(A_1 + k)}} \sum_{p=A_1+1}^{A_2} \exp\Big(\frac{1}{2}\phi''(x_0)(p - x_0)^2 \Big)$$

$$\leq (1+o(1)) \cdot \frac{\exp\{\phi_{C,k}(x_0)\}}{\sqrt{x_0(x_0+k)}} \sum_{p=-\infty}^{\infty} \exp\left(\frac{1}{2}\phi''(x_0)p^2\right)$$
$$\leq (1+o(1)) \cdot \frac{\exp\{\phi_{C,k}(x_0)\}}{\sqrt{x_0(x_0+k)}} \Big\{1 + 2\sum_{p=1}^{\infty} \exp\left(\frac{1}{2}\phi''(x_0)p^2\right)\Big\}$$
$$\leq (1+o(1)) \cdot \frac{\exp\{\phi_{C,k}(x_0)\}}{\sqrt{x_0(x_0+k)}} \Big\{1 + 2\int_0^{\infty} \exp\left(\frac{1}{2}\phi''(x_0)t^2\right)\mathrm{d}t\Big\}$$
$$\leq (1+o(1)) \cdot \frac{\exp\{\phi_{C,k}(x_0)\}}{\sqrt{x_0(x_0+k)}} \Big\{1 + \sqrt{\frac{2\pi}{-\phi''_{C,k}(x_0)}}\Big\}.$$

It is proven in the Section B.3 of the main document that $x_0(x_0+k) = C^2/4$ and $-\phi''_{C,k}(x_0) = (4/C)\sqrt{1+(k/C)^2}$. It follows, as $C \to \infty$,

$$S_2 \leq \sqrt{2\pi}(1+o(1)) \frac{\exp\{\phi_{C,k}(x_0)\}}{\sqrt{C} \cdot \left(1+(k/C)^2\right)^{1/4}}$$
$$\leq \sqrt{2\pi}(1+o(1)) \frac{\exp\{\phi_{C,k}(x_0)\}}{\sqrt{C}}.$$

Bound on S_3. Let $p > A_2$. Remark that for L large enough we also have $p > x_0 + \sqrt{x_0}$. Then, by performing two Taylor expansions, we find that there is $\bar{x} \in (x_0, x_0+\sqrt{x_0})$ and $\bar{p} \in (x_0+\sqrt{x_0}, p)$ such that

$$\begin{aligned}\phi_{C,k}(p) &= \phi_{C,k}(x_0+\sqrt{x_0}) + \phi'_{C,k}(\bar{p})(p - x_0 - \sqrt{x_0}) \\ &= \phi_{C,k}(x_0) + \frac{1}{2}\phi''_{C,k}(\bar{x})x_0 + \phi'_{C,k}(\bar{p})(p - x_0 - \sqrt{x_0}) \\ &\leq \phi_{C,k}(x_0) + \phi'_{C,k}(\bar{p})(p - x_0 - \sqrt{x_0}),\end{aligned}$$

where the last line follows because $\phi''_{C,k}(\bar{x}) < 0$. By the results of Section B.3 of the main document, we can also see that

$$\begin{aligned}\phi'_{C,k}(\bar{p}) &= \phi_{C,k}(x_0) - \log\frac{\bar{p}}{x_0} - \log\frac{\bar{p}+k}{x_0+k} \\ &= -\log\frac{\bar{p}}{x_0} - \log\frac{\bar{p}+k}{x_0+k} \\ &\leq -\log\frac{\bar{p}}{x_0} \\ &\leq -\log\left(1+\frac{1}{\sqrt{x_0}}\right) \\ &\leq -\frac{1}{1+\sqrt{x_0}}.\end{aligned}$$

Hence because $p > x_0 + \sqrt{x_0}$,

$$\phi_{C,k}(p) \leq \phi_{C,k}(x_0) - \frac{p - x_0 - \sqrt{x_0}}{1 + \sqrt{x_0}}.$$

It follows,

$$\begin{aligned} S_3 &\leq \frac{\exp\{\phi_{C,k}(x_0)\}}{\sqrt{x_0(x_0+k)}} \sum_{p > A_2} \exp\Big(- \frac{p - x_0 - \sqrt{x_0}}{1 + \sqrt{x_0}}\Big) \\ &\leq \frac{e \cdot \exp\{\phi_{C,k}(x_0)\}}{\sqrt{x_0(x_0+k)}} \cdot x_0^{-G_2} \cdot \sum_{p \geq 0} \exp\Big(- \frac{p}{1 + \sqrt{x_0}}\Big) \\ &= \frac{e \cdot \exp\{\phi_{C,k}(x_0)\}}{\sqrt{x_0(x_0+k)}} \cdot x_0^{-G_2} \cdot (1 + \sqrt{x_0}). \end{aligned}$$

It is shown in Section B.3 of the main document that $x_0(x_0 + k) = C^2/4$. Therefore, for $C \to \infty$ and G_2 sufficiently large,

$$S_3 = \frac{\exp\{\phi_{C,k}(x_0)\}}{\sqrt{C}} \cdot o(1). \qquad \square$$

S6.2. *Proof of Corollary S1.* By item (2) of Lemma S3, we obtain immediately that

$$E_L(\gamma_C, [-1,1]) \leq \sup_{x \in [-1,1]} |q_L(x) - \gamma_C(x)| \leq \sum_{k > L} |a_k(C)|. \tag{S28}$$

We let L' be the largest integer smaller than BC, for $B > 0$ large enough. Then, by the item (3) of Lemma S3, for any $D > 0$ we can choose B_0 such that for all $B > B_0$,

$$\sum_{k > L'} |a_k(C)| \leq \sum_{k > L'} e^{-Dk} \leq \frac{e^{-DBC}}{e^D - 1}.$$

By taking B, D sufficiently large, the contribution of the previous display in the rhs of (S28) is negligible. It remains to bound the sum from $L+1$ to L' (note that for B large enough, we have $L' > L$). By the item item (4) of Lemma S3, we obtain that

$$\begin{aligned} \sum_{k=L+1}^{L'} |a_k(C)| &\leq \sqrt{2\pi} \sum_{k=L+1}^{L'} \frac{\exp\{-C\varphi(k/C)\}}{\sqrt{C}} \\ &\leq \sqrt{2\pi} \int_L^{\infty} \frac{\exp\{-C\varphi(x/C)\}}{\sqrt{C}} \mathrm{d}x \end{aligned}$$

$$= \sqrt{2\pi C} \int_{L/C}^{\infty} \exp\{-C\varphi(x)\}\mathrm{d}x,$$

where the second line follows because φ is monotone increasing on (L, ∞), because $\varphi' > 0$ (see for instance Section S6.3). Interestingly, the function φ' is also monotone increasing $(L/C, \infty)$, because $\varphi'' > 0$ (see again Section S6.3). Hence, $u \geq L/C \Leftrightarrow \varphi'(u) \geq \varphi'(L/C)$, and by Markov's inequality,

$$\begin{aligned}\sum_{k=L+1}^{L'} |a_k(C)| &\leq \sqrt{\frac{2\pi}{C}} \int_{L/C}^{\infty} \frac{C\varphi'(u)\exp\{-C\varphi(u)\}}{\varphi'(L/C)}\mathrm{d}u \\ &= \sqrt{2\pi} \cdot \frac{\exp\{-C\varphi(L/C)\}}{\varphi'(L/C)\cdot\sqrt{C}}.\end{aligned}$$

Now we remark that by a Taylor expansion we have $u \in (0, L/C)$, that is $u \in (0, \zeta)$, such that $\varphi'(L/C) = \varphi'(0) + \varphi''(u)\cdot L/C = \varphi''(u)\cdot L/C$. In view of Section S6.3, we deduce that

$$\varphi'(L/C) \geq \frac{1}{\sqrt{1+\zeta^2}} \cdot \frac{L}{C},$$

and thus,

$$\sum_{k=L+1}^{L'} |a_k(C)| \leq \sqrt{2\pi(1+\zeta^2)} \cdot \frac{\sqrt{C}}{L} e^{-C\varphi(L/C)}.$$

S6.3. *Some results about the function φ.* In this section, we collect some facts about the function $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ defined in (18) of the main document. It is convenient to rewrite φ as

$$\begin{aligned}\varphi(x) := 1 - \sqrt{1+x^2} + \frac{1}{2}(-x+\sqrt{1+x^2})\log(-x+\sqrt{1+x^2}) \\ + \frac{1}{2}(x+\sqrt{1+x^2})\log(x+\sqrt{1+x^2}).\end{aligned}$$

Then,

$$\varphi'(x) = -\frac{(-x+\sqrt{1+x^2})\log(-x+\sqrt{1+x^2})}{2\sqrt{1+x^2}} + \frac{(x+\sqrt{1+x^2})\log(x+\sqrt{1+x^2})}{2\sqrt{1+x^2}},$$

$$\varphi''(x) = \frac{1}{(1+x^2)^{1/2}}, \qquad \varphi'''(x) = -\frac{x}{(1+x^2)^{3/2}}.$$

By a Taylor expansion of φ near 0, we find that there is a $y \in (0, x)$ such that

$$\varphi(x) = \varphi(0) + \varphi'(0)x + \frac{1}{2}\varphi''(0)x^2 + \frac{1}{6}\varphi'''(y)x^3 \leq \frac{x^2}{2},$$

because $\varphi(0) = \varphi'(0) = 0$ and $\varphi'''(y) \leq 0$ for all $y \geq 0$ by the computations above. Similarly, there is $y \in (0, x)$ such that,

$$|\varphi'(x)| \leq |\varphi'(0)| + |\varphi''(y)||x| \leq |x|.$$

References.

[1] Bethlehem, J.G., Keller, W.J. and Pannekoek, J. (1990). Disclosure control of microdata. *J. Amer. Statist. Assoc.* **85**, 38–45.

[2] Carota, C., Filippone, M., Leombruni, R. and Polettini, S. (2015). Bayesian nonparametric disclosure risk estimation via mixed effects log-linear models. *Ann. Appl. Statist.* **9**, 525–546.

[3] Carota, C., Filippone, M. and Polettini, S. (2018). Assessing Bayesian nonparametric log-linear models: an application to disclosure risk estimation. *Preprint: arXiv:1801.05244*

[4] Efron, B. and Morris, C (1973). Stein's estimation rule and its competitors - an empirical Bayes approach. *J. Amer. Statist. Assoc.* **68**, 117–130.

[5] Ferguson, T.S. (1973). A Bayesian Analysis of Some Nonparametric Problems. *Ann. Statist.* **1**, 209–230.

[6] Lepski, O., Nemirovski, A. and Spokoiny, V. (1999). On estimation of the L_r norm of a regression function. *Probab. Theory and Related Fields* **113**, 221–253.

[7] Luke, Y.L. (1972). Inequalities for generalized hypergeometric functions. *J. Approximation Theory*, **5**, 41–65.

[8] Manrique-Vallier, D. and Reiter, J.P. (2012). Estimating identification disclosure risk using mixed membership models. *J. Amer. Statist. Assoc.* **107** 1385–1394.

[9] Manrique-Vallier, D. and Reiter, J.P. (2014). Bayesian estimation of discrete multivariate latent structure models with structural zeros. *J. Comput. Graph. Statist.* **23** 1061–1079.

[10] Olver, F.W.J., Lozier, D.W., Boisvert, R.F. and Clark, C.W. (2010). *NIST handbook of mathematical functions*, Cambridge University Press.

[11] Robbins, H. (1956). An empirical Bayes approach to statistics. *Proc. 3rd Berkeley Symp.*,**1**, 157–163.

[12] Samuels, S.M. (1998). A Bayesian, species-sampling-inspired approach to the uniques problem in microdata disclosure risk assessment. *J. Off. Statist.* **14**, 373–383.

[13] Skinner, C., Marsh, C., Openshaw, S. and Wymer, C. (1994). Disclosure control for census microdata. *J. Off. Stat.* **10**, 31–51.

[14] Tsybakov, A. B. (2009) *Introduction to nonparametric estimation.* Springer Science & Business Media.

[15] Wu, Y. and Yang, P. (2016). Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Trans. Inform. Theory* **62**, 3702–3720.

[16] Wu, Y. and Yang, P. (2015). Chebyshev polynomials, moment matching, and optimal estimation of the unseen. *Ann. Statist.*, to appear.

Federico Camerlenghi
Department of Economics, Management and Statistics
University of Milano - Bicocca
20126 Milano, Italy.
E-mail: federico.camerlenghi@unimib.it

Stefano Favaro
Department of Economics and Statistics
Univeristy of Torino
10134 Torino, Italy.
E-mail: stefano.favaro@unito.it

Zacharie Naulet
Université Paris-Saclay
CNRS, Laboratoire de mathématiques d'Orsay
91405 Orsay, France.
E-mail: zacharie.naulet@math.u-psud.fr

Francesca Panero
Department of Statistics
University of Oxford
OX1 3LB Oxford, United Kingdom.
E-mail: francesca.panero@stats.ox.ac.uk

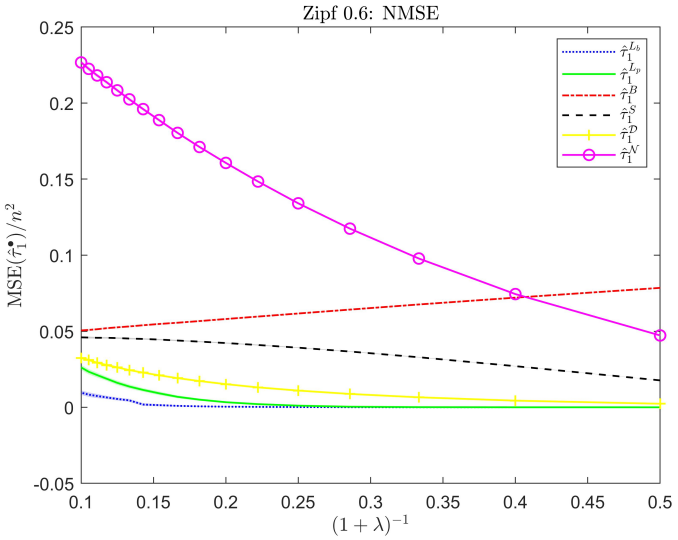


FIG S1. *The normalized mean squared error as a function of the sampling fraction* $(1+\lambda)^{-1}$ *when the distribution of the cell's probabilities is a Zipf with parameter* $s = 0.6$. *Each curve corresponds to a different estimator of* τ_1*: i) the nonparametric estimator with Binomial smoothing* $\hat{\tau}_1^{L_b}$*; ii) the nonparametric estimator with Poisson smoothing* $\hat{\tau}_1^{L_p}$*; iii) the naive nonparametric estimator* $\hat{\tau}_1^{\mathcal{N}}$*; iv) the Bayesian nonparametric estimator* $\hat{\tau}_1^{\mathcal{D}}$*; v) the parametric empirical Bayes estimator* $\hat{\tau}_1^{B}$*; vi) the parametric empirical Bayes estimator* $\hat{\tau}_1^{S}$. *The shaded bands corresponds to one standard deviation.*

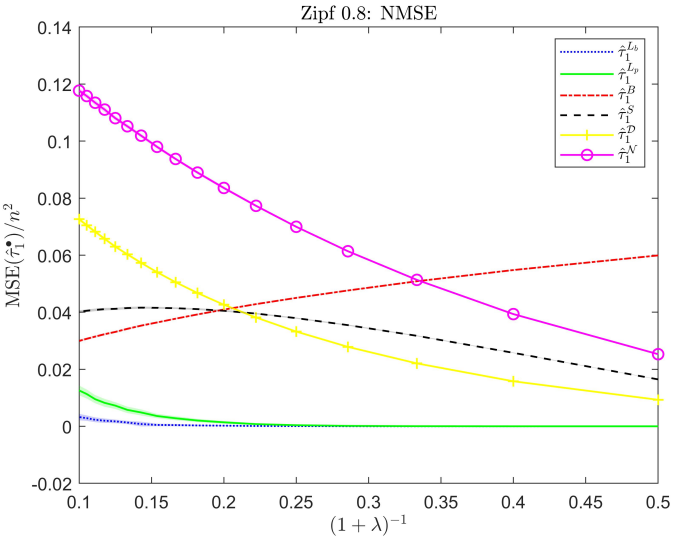


FIG S2. *The normalized mean squared error as a function of the sampling fraction* $(1+\lambda)^{-1}$ *when the distribution of the cell's probabilities is a Zipf with parameter* $s = 0.8$. *Each curve corresponds to a different estimator of* τ_1*: i) the nonparametric estimator with Binomial smoothing* $\hat{\tau}_1^{L_b}$*; ii) the nonparametric estimator with Poisson smoothing* $\hat{\tau}_1^{L_p}$*; iii) the naive nonparametric estimator* $\hat{\tau}_1^{\mathcal{N}}$*; iv) the Bayesian nonparametric estimator* $\hat{\tau}_1^{\mathcal{D}}$*; v) the parametric empirical Bayes estimator* $\hat{\tau}_1^{B}$*; vi) the parametric empirical Bayes estimator* $\hat{\tau}_1^{S}$. *The shaded bands corresponds to one standard deviation.*

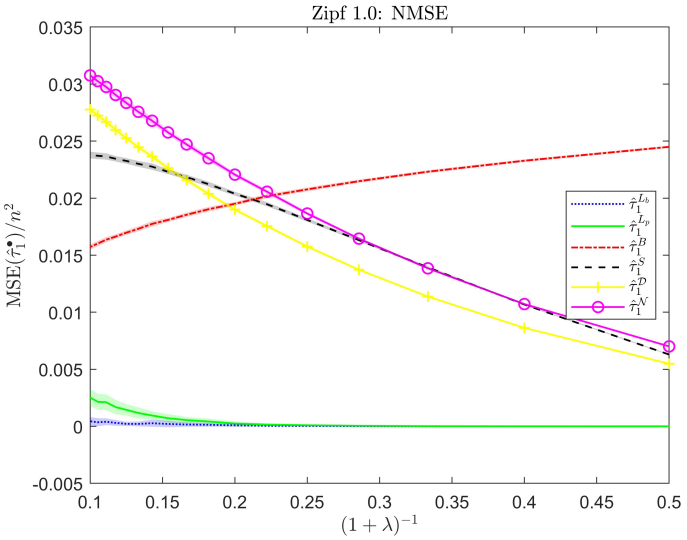


FIG S3. *The normalized mean squared error as a function of the sampling fraction* $(1+\lambda)^{-1}$ *when the distribution of the cell's probabilities is a Zipf with parameter* $s = 1.0$. *Each curve corresponds to a different estimator of* τ_1*: i) the nonparametric estimator with Binomial smoothing* $\hat{\tau}_1^{L_b}$*; ii) the nonparametric estimator with Poisson smoothing* $\hat{\tau}_1^{L_p}$*; iii) the naive nonparametric estimator* $\hat{\tau}_1^{\mathcal{N}}$*; iv) the Bayesian nonparametric estimator* $\hat{\tau}_1^{\mathcal{D}}$*; v) the parametric empirical Bayes estimator* $\hat{\tau}_1^{B}$*; vi) the parametric empirical Bayes estimator* $\hat{\tau}_1^{S}$. *The shaded bands corresponds to one standard deviation.*

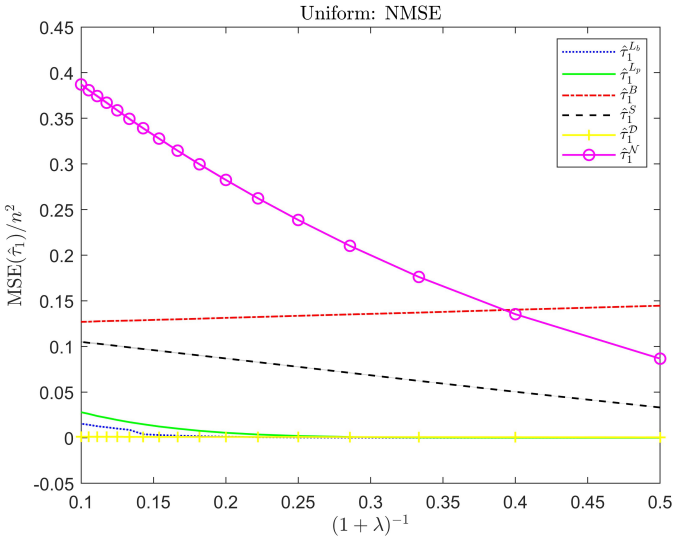


FIG S4. *The normalized mean squared error as a function of the sampling fraction* $(1+\lambda)^{-1}$ *when the cell's probabilities are uniform distributed. Each curve corresponds to a different estimator of* τ_1*: i) the nonparametric estimator with Binomial smoothing* $\hat{\tau}_1^{L_b}$*; ii) the nonparametric estimator with Poisson smoothing* $\hat{\tau}_1^{L_p}$*; iii) the naive nonparametric estimator* $\hat{\tau}_1^{\mathcal{N}}$*; iv) the Bayesian nonparametric estimator* $\hat{\tau}_1^{\mathcal{D}}$*; v) the parametric empirical Bayes estimator* $\hat{\tau}_1^{B}$*; vi) the parametric empirical Bayes estimator* $\hat{\tau}_1^{S}$*. The shaded bands corresponds to one standard deviation.*

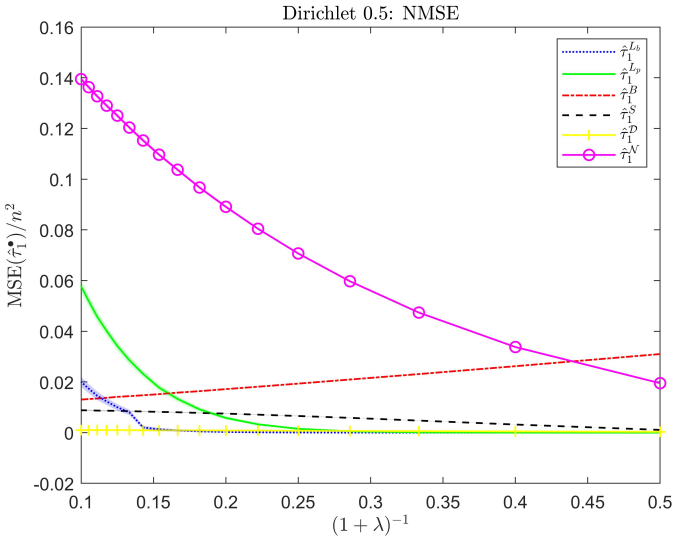


FIG S5. *The normalized mean squared error as a function of the sampling fraction* $(1+\lambda)^{-1}$ *when the distribution of the cell's probabilities is a uniform Dirichlet distribution with respective parameter* $\beta = 0.5$. *Each curve corresponds to a different estimator of* τ_1*: i) the nonparametric estimator with Binomial smoothing* $\hat{\tau}_1^{L_b}$*; ii) the nonparametric estimator with Poisson smoothing* $\hat{\tau}_1^{L_p}$*; iii) the naive nonparametric estimator* $\hat{\tau}_1^{\mathcal{N}}$*; iv) the Bayesian nonparametric estimator* $\hat{\tau}_1^{\mathcal{D}}$*; v) the parametric empirical Bayes estimator* $\hat{\tau}_1^{B}$*; vi) the parametric empirical Bayes estimator* $\hat{\tau}_1^{S}$. *The shaded bands corresponds to one standard deviation.*

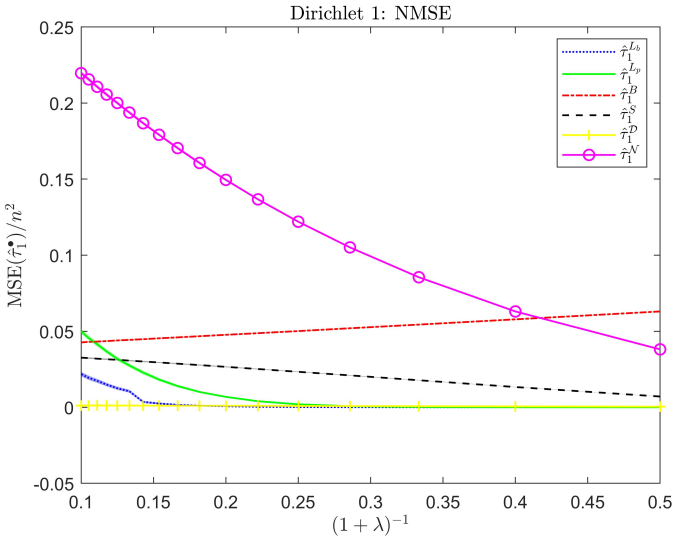


FIG S6. *The normalized mean squared error as a function of the sampling fraction* $(1+\lambda)^{-1}$ *when the distribution of the cell's probabilities is a uniform Dirichlet distribution with respective parameter* $\beta = 1.0$. *Each curve corresponds to a different estimator of* τ_1*: i) the nonparametric estimator with Binomial smoothing* $\hat{\tau}_1^{L_b}$*; ii) the nonparametric estimator with Poisson smoothing* $\hat{\tau}_1^{L_p}$*; iii) the naive nonparametric estimator* $\hat{\tau}_1^{\mathscr{N}}$*; iv) the Bayesian nonparametric estimator* $\hat{\tau}_1^{\mathscr{D}}$*; v) the parametric empirical Bayes estimator* $\hat{\tau}_1^{B}$*; vi) the parametric empirical Bayes estimator* $\hat{\tau}_1^{S}$. *The shaded bands corresponds to one standard deviation.*

	Zipf 0.6	Zipf 0.8	Zipf 1
True τ_1	112780	82254	42397
$\hat{\tau}_1^{L_b}$	$116533 \in (115361, 117704)$	$84478 \in (83041, 85916)$	$43370 \in (41980, 44760)$
$\hat{\tau}_1^{L_p}$	$124242 \in (123380, 125104)$	$89443 \in (88195, 90690)$	$45307 \in (44188, 46427)$
$\hat{\tau}_1^{\mathscr{N}}$	$32623 \in (32580, 32666)$	$24436 \in (24386, 24485)$	$12593 \in (12555, 12630)$
$\hat{\tau}_1^{\mathscr{D}}$	$88030 \in (87699, 88362)$	$40983 \in (40833, 41133)$	$14740 \in (14688, 14792)$
$\hat{\tau}_1^{B}$	$64587 \in (64525, 64650)$	$41815 \in (41714, 41915)$	$14362 \in (14312, 14412)$
$\hat{\tau}_1^{S}$	$71651 \in (71543, 71759)$	$42022 \in (41900, 42145)$	$13738 \in (13690, 13787)$

	Uniform	Dirichlet 0.5	Dirichlet 1
True τ_1	143375	92849	112468
$\hat{\tau}_1^{L_b}$	$149823 \in (149127, 150520)$	$95806 \in (94658, 96955)$	$117449 \in (116465, 118433)$
$\hat{\tau}_1^{L_p}$	$157967 \in (157408, 158526)$	$108040 \in (107174, 108907)$	$128879 \in (128150, 129607)$
$\hat{\tau}_1^{\mathscr{N}}$	$37424 \in (37392, 37457)$	$33133 \in (33086, 33181)$	$35147 \in (35110, 35184)$
$\hat{\tau}_1^{\mathscr{D}}$	$149121 \in (148619, 149623)$	$98586 \in (98178, 98993)$	$118696 \in (118285, 119106)$
$\hat{\tau}_1^{B}$	$71141 \in (71110, 71172)$	$66620 \in (66568, 66672)$	$68820 \in (68782, 68858)$
$\hat{\tau}_1^{S}$	$84631 \in (84565, 84697)$	$75504 \in (75404, 75604)$	$79853 \in (79776, 79930)$

TABLE S1

Estimation of τ_1 for several simulated scenarios, when the size of the population is $\bar{n} = 10^6$ and $(\lambda + 1)^{-1} = 1/5$. Each column corresponds to a different choice of the distribution over the cells' probabilities. The first line displays the true value of τ_1, while the other rows contain the estimates and the empirical bands based on one standard deviation. All the experiments are averaged over 100 *iterations.*

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

Title of Paper	**Optimal disclosure risk assessment**
Publication Status	☑ Published □ Accepted for Publication □ Submitted for Publication □ Unpublished and unsubmitted work written in a manuscript style
Publication Details	Federico Camerlenghi, Stefano Favaro, Zacharie Naulet, Francesca Panero "Optimal disclosure risk assessment", The Annals of Statistics, Ann. Statist. 49(2), 723-744, (April 2021)

Student Confirmation

Student Name:	Francesca Panero		
Contribution to the Paper	All the authors contributed equally to the manuscript. I derived the estimator under the empirical Bayes approach (Appendix A.2) and the study of its expectation and variance (Theorem 1 and Corollary 1). I derived the smoothed estimator and its characteristics (Theorem 2, Proposition 1, Proposition 2). I have contributed to the design of the experiments in Supplementary S3.		
Signature 		Date	19/04/2022

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title:	Professor François Caron		
Supervisor comments			
Signature 		Date	19/04/2022

This completed form should be included in the thesis, at the end of the relevant chapter.

Chapter 6

Discussion

This final chapter contains the summary of the work presented in this thesis and suggestions for possible research directions.

6.1 Summary

This thesis contains four pieces of work regarding two statistical modelling challenges: networks and disclosure risk assessment.

In chapter 2, I presented a novel methodology to describe sparse and dense spatial networks. This random graph model relies on completely random measures as prior distributions for the variables of the model to achieve desirable network properties such as sparsity, power-law degree distributions for low degree nodes and positive clustering. We describe how to sample from the model in sub-quadratic time. We provide a Markov Chain Monte Carlo algorithm to sample from an approximate posterior distribution of the model's parameters and variables and show that the algorithm works well on simulated data. The Bayesian approach allows conveniently to quantify the uncertainty in the estimations.

Chapter 3 deals with the asymptotic studies on graphs generated in the graphex process framework, under some regular variation assumptions. We provide

the relation between number of nodes and edges and show how the graphex process can describe both sparse and dense networks. We show under which assumptions it is possible to obtain a power-law degree distribution for low degree nodes and a double power-law with different exponents for high and low degrees. We provide the asymptotic limits for the global and average local clustering coefficients and prove central limit theorems for the number of nodes and subgraph counts. To show some practical cases, we apply these results to numerous sparse graphon functions. Finally, we extend some of the findings to a framework enhanced by local properties, for example spatial structures or communities.

Chapter 4 presents a study of statistical quantification of disclosure risk in terms of the number of individuals whose answers in a microdata file make them uniques in the sample and in the underlying population. The estimation is based on a prior distribution commonly used in Bayesian nonparametric, the Pitman–Yor process, from which the estimator inherits a convenient closed-form representation of its posterior distribution, a scheme to sample efficiently from it and quantify the uncertainty of the estimate. We show empirically on simulated and real data that the estimator is particularly well suited to describe populations generated under distributions with power-law or geometric tails.

Chapter 5 is chronologically the first work I did on statistical estimation of disclosure risk. The work was motivated by the desire to find an estimator for the number of sample uniques that are also population uniques without assuming any shape on the distribution of the population. Our proposal is fully nonparametric and is very easy to understand and compute. We prove that the estimator is optimal since, under an assumption on the relative sample size, it has vanishing normalised mean squared error with matching upper and lower bounds, and when this assumption is violated it is impossible to find a nonparametric estimator with guarantees of vanishing error. This study motivated the need of a more specific assumption on the distribution of the population, which subsequently gave birth to the idea for chapter 4.

6.2 Extentions

The papers and manuscripts I have presented open some questions that require further investigations and could lead to new research directions. I will explore some of them here.

6.2.1 Efficient inference for sparse spatial random graphs

Chapter 2 provides an inference scheme, reliant on MCMC techniques, to sample from an approximation of the posterior distribution of the variables and parameters of the model. The time complexity of such algorithm is squared in the number of nodes and therefore cannot be run in a reasonable time on datasets with more than a few thousands of nodes. To insure a wider applicability of the proposed model, a faster algorithm is required. Inspired by the work of [Rastelli et al., 2018], an accessible possibility would be to find an approximation of the likelihood function which relies on partitions of the space domain and of the sociability layers, as already done in the sampling scheme proposed in chapter 2. Another possibility, more challenging yet more elegant and leading to an exact algorithm, would be to use Poisson-minibatching methods, as proposed in [Zhang and De Sa, 2019]. This technique requires to find new auxiliary Poisson variables whose presence induces an augmented joint likelihood with lower computational complexity.
As the influence of space on the connection function increases, so do the challenges brought by the multimodal posterior. The second direction to develop a more efficient algorithm is that of parallel tempering techniques ([Swendsen and Wang, 1986]), which would encourage the Markov chains to explore different regions of the space and not to get stuck in local optima.

6.2.2 Small worldness and spatial asymptotics

An interesting property of real networks is that it is possible to traverse the network with a surprisingly low number of steps, quantified as proportional to the logarithm of the number of nodes. Scale-free networks usually induce this

behaviour, and it would be exciting to understand if it holds true in the case of the sparse spatial random graphs and how the distance changes as a function of γ. From similar studies in [Deprez and Wüthrich, 2018], I expect small and ultra small world effects to hold, depending on the relation between the parameter tuning the effects of the distance γ, the exponent of the regularly varying distribution of the sociability weights α and the dimension of the space of locations d. Similarly, it could be interesting to explore the percolation properties of such model to understand the emergence of a giant component.

The asymptotic properties of networks generated under the graphex framework proved in chapter 2 and chapter 3 have always been studied as functions of the time of appearance of the nodes t. No less important would be to explore the world of asymptotics in space, as $x_{\max}$ tends to infinity. This would also allow to compare more easily the sparse spatial random graph with other spatial network models in the literature.

6.2.3 Applications of spatial network models

The sparse spatial random graph model of chapter 2 has been applied so far only to simulated data, while testing it on real world data remains to be explored. I would start from the US airport dataset illustrated in fig. 1.1. Airports have real locations to test against and therefore it would be easy to understand if the model is able to elicit spatial information by studying the correlation between the inferred locations and the longitude and latitude of the nodes. A space of dimension $d = 2$ and a geodesic distance function seem to be the most natural choices to describe such network, but it would be interesting to understand if different configurations could reveal other insights. Furthermore, [Li and Cai, 2004] and [Paleari et al., 2010] observed that some airports networks display a double power-law degree distribution with different exponents for large and small degree nodes. The use of CRMs such as the generalised gamma Pareto process (as constructed in [Ayed et al., 2019]) as prior for the sociability weights could induce such behaviour and might provide a better fit of the degree distribution.

Many biological applications can be described by networks which are influenced by their spatial structure. An example of this is neuroscience. The network of connectomes, white matter fibers in the brain, are determined by a mixture of their concrete spatial coordinates and an additional latent structure (for example, see [Aliverti and Durante, 2019]). The degree distribution of such networks seem to be better described by the Weibull distribution or a power-law with exponential cut-off, instead of pure power-laws ([Gastner and Ódor, 2016]). It would be interesting to study the performance of the sparse spatial random graph model on brain networks characterised by the power-law with exponential cut-off. More challenging, but possibly more interesting, would be the opportunity to modify the model to accommodate different types of degree distributions.

6.2.4 Disclosure risk assessment in presence of structural zeros

In the setting of disclosure risk assessment, a question of modelling interest is that of structural zeros. Structural zeros are defined as combinations of individual records that are impossible to observe (for example, an 8 years old with kids). Since they represent impossible combinations, it would be wise to account for them in the modelling scheme. A possibility is to employ spike and slab priors that mix two probability distributions, the first being that chosen for the observable combinations (for example, the Pitman–Yor process in chapter 4) and the second being the Dirac measure at 0 for structural zeros. Examples of such priors can be found in [Scarpa and Dunson, 2009] and [Canale et al., 2017].

Bibliography

[Aalen, 1992] Aalen, O. (1992). Modelling heterogeneity in survival analysis by the compound Poisson distribution. *The Annals of Applied Probability*, pages 951–972.

[Airoldi et al., 2008] Airoldi, E. M., Blei, D., Fienberg, S. E., and Xing, E. (2008). Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9:1981–2014.

[Albert et al., 1999] Albert, R., Jeong, H., and Barabási, A.-L. (1999). Diameter of the world-wide web. *nature*, 401(6749):130–131.

[Aldous, 1985a] Aldous, D. (1985a). Exchangeability and related topics. In *Ecole d'été de Probabilités de Saint-Flour XIII - 1983*, pages 1–198. Springer.

[Aldous, 1981] Aldous, D. J. (1981). Representations for partially exchangeable arrays of random variables. *Journal of Multivariate Analysis*, 11(4):581–598.

[Aldous, 1985b] Aldous, D. J. (1985b). Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII—1983*, pages 1–198. Springer.

[Aldous, 2008] Aldous, D. J. (2008). Spatial transportation networks with transfer costs: asymptotic optimality of hub-and-spoke models. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 145, pages 471–487. Cambridge University Press.

[Aliverti and Durante, 2019] Aliverti, E. and Durante, D. (2019). Spatial modeling of brain connectivity data via latent distance models with nodes

clustering. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 12(3):185–196.

[Ayed et al., 2019] Ayed, F., Lee, J., and Caron, F. (2019). Beyond the chinese restaurant and pitman-yor processes: Statistical models with double power-law behavior. *arXiv preprint arXiv:1902.04714*.

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

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

[Bayardo and Agrawal, 2005] Bayardo, R. J. and Agrawal, R. (2005). Data privacy through optimal k-anonymization. In *21st International conference on data engineering (ICDE'05)*, pages 217–228. IEEE.

[Beirlant et al., 2004] Beirlant, J., Goegebeur, Y., Segers, J., and Teugels, J. L. (2004). *Statistics of extremes: theory and applications*, volume 558. John Wiley & Sons.

[Berger et al., 2014] Berger, N., Borgs, C., Chayes, J. T., and Saberi, A. (2014). Asymptotic behavior and distributional limits of preferential attachment graphs. *The Annals of Probability*, 42(1):1–40.

[Bertoin, 2006] Bertoin, J. (2006). *Random fragmentation and coagulation processes.* Cambridge Studies in Advanced Mathematics (Book 102). Cambridge University Press.

[Bethlehem et al., 1990] Bethlehem, J. G., Keller, W. J., and Pannekoek, J. (1990). Disclosure control of microdata. *Journal of the American Statistical Association*, 85(409):38–45.

[Bingham et al., 1987] Bingham, N. H., Goldie, C. M., and Teugels, J. L. (1987). *Regular variation*, volume 27. Cambridge university press.

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

[Bollobás and Riordan, 2009] Bollobás, B. and Riordan, O. (2009). Metrics for sparse graphs. In Huczynska, S., Mitchell, J., and Roney-Dougal, C., editors, *Surveys in combinatorics*, volume 365 of *London Mathematical Society Lecture Note Series*, pages 211–287. Cambridge University Press, arXiv:0708.1919.

[Borgs et al., 2018] Borgs, C., Chayes, J., Cohn, H., and Zhao, Y. (2018). An L^p theory of sparse graph convergence II: LD convergence, quotients, and right convergence. *The Annals of Probability*, 46(1):337–396.

[Borgs et al., 2019] Borgs, C., Chayes, J., Cohn, H., and Zhao, Y. (2019). An L^p theory of sparse graph convergence i: Limits, sparse random graph models, and power law distributions. *Transactions of the American Mathematical Society*, 372(5):3019–3062.

[Borgs et al., 2017] Borgs, C., Chayes, J., and Gamarnik, D. (2017). Convergent sequences of sparse graphs: A large deviations approach. *Random Structures & Algorithms*, 51(1):52–89.

[Borgs et al., 2010] Borgs, C., Chayes, J. T., and Lovász, L. (2010). Moments of two-variable functions and the uniqueness of graph limits. *Geometric And Functional Analysis*, 19(6):1597–1619.

[Brix, 1999] Brix, A. (1999). Generalized gamma measures and shot-noise Cox processes. *Advances in Applied Probability*, 31(4):929–953.

[Broderick and Cai, 2016] Broderick, T. and Cai, D. (2016). Edge-exchangeable graphs and sparsity. *arXiv preprint arXiv:1603.06898*.

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

[Caldarelli, 2007] Caldarelli, G. (2007). *Scale-free networks: complex webs in nature and technology*. Oxford University Press.

[Canale et al., 2017] Canale, A., Lijoi, A., Nipoti, B., and Prünster, I. (2017). On the pitman–yor process with spike and slab base measure. *Biometrika*, 104(3):681–697.

[Caron and Fox, 2017] Caron, F. and Fox, E. (2017). Sparse graphs using exchangeable random measures. *Journal of the Royal Statistical Society B*, 79:1–44. Part 5.

[Carota et al., 2015] Carota, C., Filippone, M., Leombruni, R., and Polettini, S. (2015). Bayesian nonparametric disclosure risk estimation via mixed effects log-linear models. *The Annals of Applied Statistics*, 9(1):525–546.

[Cerquetti, 2013] Cerquetti, A. (2013). Bayesian nonparametric estimation of global disclosure risk. In *9th Scientific Meeting of the CLAssification and Data Analysis Group, Italian Statistical Society*.

[Chatterjee and Diaconis, 2013] Chatterjee, S. and Diaconis, P. (2013). Estimating and understanding exponential random graph models. *The Annals of Statistics*, 41(5):2428–2461.

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

[Clauset et al., 2009] Clauset, A., Shalizi, C. R., and Newman, M. E. J. (2009). Power-law distributions in empirical data. *SIAM review*, 51(4):661–703.

[Crane and Dempsey, 2015] Crane, H. and Dempsey, W. (2015). A framework for statistical network modeling. *arXiv preprint arXiv:1509.08185*.

[Daley and Smith, 2013] Daley, T. and Smith, A. D. (2013). Predicting the molecular complexity of sequencing libraries. *Nature methods*, 10(4):325–327.

[Danos, 1998] Danos, M. (1998). Fractals and chaos in geology and geophysics.

[de Finetti, 1931] de Finetti, B. (1931). Funzione caratteristica di un fenomeno aleatorio. *Atti della R. Academia Nazionale dei Lincei, Serie 6. Memorie, Classe di Scienze Fisiche, Mathematice e Naturale*, 4:251–299.

[De Finetti, 1937] De Finetti, B. (1937). La prévision: ses lois logiques, ses sources subjectives. In *Annales de l'institut Henri Poincaré*, volume 7, pages 1–68.

[de Solla Price, 1965] de Solla Price, D. J. (1965). Networks of scientific papers. *Science*, 149(3683):510–515.

[Deprez and Wüthrich, 2018] Deprez, P. and Wüthrich, M. V. (2018). Scale-free percolation in continuum space. *Communications in Mathematics and Statistics*, 7(3):269–308.

[Devroye, 2009] Devroye, L. (2009). Random variate generation for exponentially and polynomially tilted stable distributions. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 19(4):18.

[Diaconis and Janson, 2008] Diaconis, P. and Janson, S. (2008). Graph limits and exchangeable random graphs. *Rendiconti di Matematica e delle sue Applicazioni. Serie VII*, pages 33–61.

[Ding et al., 2018] Ding, Z., Wang, Y., Wang, G., Zhang, D., and Kifer, D. (2018). Detecting violations of differential privacy. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, pages 475–489.

[Duncan and Lambert, 1989] Duncan, G. and Lambert, D. (1989). The risk of disclosure for microdata. *Journal of Business & Economic Statistics*, 7(2):207–217.

[Dwork, 2006] Dwork, C. (2006). Differential privacy. In *33rd International Colloquium on Automata, Languages and Programming, part II (ICALP 2006)*, volume 4052 of *Lecture Notes in Computer Science*, pages 1–12. Springer Verlag.

[Dwork et al., 2019] Dwork, C., Kohli, N., and Mulligan, D. (2019). Differential privacy in practice: Expose your epsilons! *Journal of Privacy and Confidentiality*, 9(2).

[Dwork et al., 2006] Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer.

[Efron and Thisted, 1976] Efron, B. and Thisted, R. (1976). Estimating the number of unseen species: How many words did shakespeare know? *Biometrika*, 63(3):435–447.

[Erdös and Rényi, 1959] Erdös, P. and Rényi, A. (1959). On random graphs. *Publicationes Mathematicae*, 6:290–297.

[Faloutsos et al., 1999] Faloutsos, M., Faloutsos, P., and Faloutsos, C. (1999). On power-law relationships of the internet topology. *ACM SIGCOMM computer communication review*, 29(4):251–262.

[Ferguson, 1973] Ferguson, T. S. (1973). A bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230.

[Fisher et al., 1943] Fisher, R. A., Corbet, A. S., and Williams, C. B. (1943). The relation between the number of species and the number of individuals in a random sample of an animal population. *The Journal of Animal Ecology*, pages 42–58.

[Florencio and Herley, 2007] Florencio, D. and Herley, C. (2007). A large-scale study of web password habits. In *Proceedings of the 16th international conference on World Wide Web*, pages 657–666.

[Gabaix, 1999] Gabaix, X. (1999). Zipf's law for cities: an explanation. *The Quarterly journal of economics*, 114(3):739–767.

[Gao et al., 2007] Gao, Z., Tseng, C.-h., Pei, Z., and Blaser, M. J. (2007). Molecular analysis of human forearm superficial skin bacterial biota. *Proceedings of the National Academy of Sciences*, 104(8):2927–2932.

[Garfinkel et al., 2018] Garfinkel, S. L., Abowd, J. M., and Powazek, S. (2018). Issues encountered deploying differential privacy. In *Proceedings of the 2018 Workshop on Privacy in the Electronic Society*, pages 133–137.

[Gastner and Newman, 2006] Gastner, M. T. and Newman, M. E. (2006). Optimal design of spatial distribution networks. *Physical Review E*, 74(1):016117.

[Gastner and Ódor, 2016] Gastner, M. T. and Ódor, G. (2016). The topology of large open connectome networks for the human brain. *Scientific reports*, 6(1):1–11.

[Gelman et al., 2003] Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian data analysis (second edition)*. Chapman and Hall/CRC.

[Ghosal and Van der Vaart, 2017] Ghosal, S. and Van der Vaart, A. (2017). *Fundamentals of nonparametric Bayesian inference*, volume 44. Cambridge University Press.

[Gnedin et al., 2007] Gnedin, A., Hansen, B., and Pitman, J. (2007). Notes on the occupancy problem with infinitely many boxes: general asymptotics and power laws. *Probability surveys*, 4:146–171.

[Golosovsky, 2017] Golosovsky, M. (2017). Power-law citation distributions are not scale-free. *Physical Review E*, 96(3):032306.

[Good, 1953] Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3-4):237–264.

[Good and Toulmin, 1956] Good, I. J. and Toulmin, G. H. (1956). The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika*, 43(1-2):45–63.

[Goodman, 1949] Goodman, L. A. (1949). On the estimation of the number of classes in a population. *The Annals of Mathematical Statistics*, 20(4):572–579.

[Harris, 2013] Harris, J. K. (2013). *An introduction to exponential random graph modeling*, volume 173. Sage Publications.

[Hjort et al., 2010] Hjort, N. L., Holmes, C., Müller, P., and Walker, S. G. (2010). *Bayesian nonparametrics*, volume 28. Cambridge University Press.

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

[Holland and Leinhardt, 1981] Holland, P. W. and Leinhardt, S. (1981). An exponential family of probability distributions for directed graphs. *Journal of the american Statistical association*, 76(373):33–50.

[Hoover, 1979] Hoover, D. N. (1979). Relations on probability spaces and arrays of random variables. *Preprint, Institute for Advanced Study, Princeton, NJ*.

[Hougaard, 1986] Hougaard, P. (1986). Survival models for heterogeneous populations derived from stable distributions. *Biometrika*, 73(2):387–396.

[Ionita-Laza et al., 2009] Ionita-Laza, I., Lange, C., and Laird, N. M. (2009). Estimating the number of unseen variants in the human genome. *Proceedings of the National Academy of Sciences*, 106(13):5008–5013.

[Janson, 2016] Janson, S. (2016). Graphons and cut metric on sigma-finite measure spaces. *arXiv:1608.01833*.

[Janson, 2017] Janson, S. (2017). Discussion on "sparse graphs using exchangeable random measures". *Journal of the Royal Statistical Society B*, 79:1–44. Part 5.

[Kallenberg, 1990] Kallenberg, O. (1990). Exchangeable random measures in the plane. *Journal of Theoretical Probability*, 3(1):81–136.

[Karrer and Newman, 2011] Karrer, B. and Newman, M. E. (2011). Stochastic blockmodels and community structure in networks. *Physical Review E*, 83(1):016107.

[Kenig and Tassa, 2012] Kenig, B. and Tassa, T. (2012). A practical approximation algorithm for optimal k-anonymity. *Data Mining and Knowledge Discovery*, 25(1):134–168.

[Kenny et al., 2021] Kenny, C. T., Kuriwaki, S., McCartan, C., Rosenman, E. T., Simko, T., and Imai, K. (2021). The use of differential privacy for census data and its impact on redistricting: The case of the 2020 us census. *Science advances*, 7(41):eabk3283.

[Kingman, 1967] Kingman, J. F. C. (1967). Completely random measures. *Pacific Journal of Mathematics*, 21(1):59–78.

[Kingman, 1993] Kingman, J. F. C. (1993). *Poisson processes*, volume 3. Oxford University Press, USA.

[Kryven, 2017] Kryven, I. (2017). General expression for the component size distribution in infinite configuration networks. *Physical Review E*, 95(5):052303.

[Lee and Whitmore, 1993] Lee, M.-L. T. and Whitmore, G. A. (1993). Stochastic processes directed by randomized time. *Journal of applied probability*, pages 302–314.

[Lee, 2019] Lee, J., M. X. C. F. (2019). A unified construction for series representations and finite approximations of completely random measures.

[Li and Cai, 2004] Li, W. and Cai, X. (2004). Statistical analysis of airport network of china. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 69:046106.

[Lijoi et al., 2005] Lijoi, A., Mena, R. H., and Prünster, I. (2005). Bayesian nonparametric analysis for a generalized dirichlet process prior. *Statistical Inference for Stochastic Processes*, 8(3):283–309.

[Liu and Zhao, 2012] Liu, M. and Zhao, L. (2012). An integrated and dynamic optimisation model for the multi-level emergency logistics network in anti-bioterrorism system. *International Journal of Systems Science*, 43(8):1464–1478.

[Lovász and Szegedy, 2006] Lovász, L. and Szegedy, B. (2006). Limits of dense graph sequences. *Journal of Combinatorial Theory, Series B*, 96(6):933–957.

[Lusher et al., 2013] Lusher, D., Koskinen, J., and Robins, G. (2013). *Exponential random graph models for social networks: Theory, methods, and applications*, volume 35. Cambridge University Press.

[Martín and Goldenfeld, 2006] Martín, H. G. and Goldenfeld, N. (2006). On the origin and robustness of power-law species–area relationships in ecology. *Proceedings of the National Academy of Sciences*, 103(27):10310–10315.

[Matthews and Harel, 2011] Matthews, G. J. and Harel, O. (2011). Data confidentiality: A review of methods for statistical disclosure limitation and methods for assessing privacy. *Statistics Surveys*, 5:1–29.

[Meyerson and Williams, 2004] Meyerson, A. and Williams, R. (2004). On the complexity of optimal k-anonymity. In *Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 223–228.

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

[Mueller and Santos-Lozada, 2022] Mueller, J. T. and Santos-Lozada, A. R. (2022). The 2020 us census differential privacy method introduces disproportionate discrepancies for rural and non-white populations. *Population Research and Policy Review*, pages 1–14.

[Newman, 2010] Newman, M. E. J. (2010). *Networks: An Introduction*. Oxford University Press.

[Orbanz and Roy, 2015] Orbanz, P. and Roy, D. M. (2015). Bayesian models of graphs, arrays and other exchangeable random structures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):437–461.

[Orlitsky et al., 2016] Orlitsky, A., Suresh, A. T., and Wu, Y. (2016). Optimal prediction of the number of unseen species. *Proceedings of the National Academy of Sciences*, 113(47):13283–13288.

[Paleari et al., 2010] Paleari, S., Redondi, R., and Malighetti, P. (2010). A comparative study of airport connectivity in China, Europe and US: Which network provides the best service to passengers? *Transportation Research Part E: Logistics and Transportation Review*, 46(2):198–210.

[Pitman and Yor, 1997] Pitman, J. and Yor, M. (1997). The two-parameter poisson-dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, pages 855–900.

[Powers, 1998] Powers, D. M. W. (1998). Applications and explanations of Zipf's law. In *New Methods in Language Processing and Computational Natural Language Learning*.

[Rastelli et al., 2018] Rastelli, R., Maire, F., and Friel, N. (2018). Computationally efficient inference for latent position network models. *arXiv preprint arXiv:1804.02274*.

[Resnick, 1987] Resnick, S. (1987). *Extreme values, point processes and regular variation*. Springer-Verlag, New York.

[Rinott and Shlomo, 2006] Rinott, Y. and Shlomo, N. (2006). A generalized negative binomial smoothing model for sample disclosure risk estimation. In *International Conference on Privacy in Statistical Databases*, pages 82–93. Springer.

[Robbins, 1956] Robbins, H. E. (1956). An empirical bayes approach to statistics.

[Robert et al., 2007] Robert, C. P. et al. (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation*, volume 2. Springer.

[Samarati and Sweeney, 1998] Samarati, P. and Sweeney, L. (1998). Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression.

[Samuels, 1998] Samuels, S. M. (1998). A bayesian, species-sampling-inspired approach to the uniques problem in microdata disclosure risk assessment. *Journal of Official Statistics*, 14(4):373.

[Scarpa and Dunson, 2009] Scarpa, B. and Dunson, D. B. (2009). Bayesian hierarchical functional data analysis via contaminated informative priors. *Biometrics*, 65(3):772–780.

[Scholkmann, 2016] Scholkmann, F. (2016). Power-law scaling of the impact crater size-frequency distribution on pluto: A preliminary analysis based on first images from new horizons flyby. *Progress in Physics*, 12(1):26–29.

[Simon, 1955] Simon, H. A. (1955). On a class of skew distribution functions. *Biometrika*, 42(3/4):425–440.

[Skinner and Elliot, 2002a] Skinner, C. J. and Elliot, M. J. (2002a). A measure of disclosure risk for microdata. *Journal of the Royal Statistical Society: series B (statistical methodology)*, 64(4):855–867.

[Skinner and Elliot, 2002b] Skinner, C. J. and Elliot, M. J. (2002b). A measure of disclosure risk for microdata. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 64(4):855–867.

[Solé et al., 2003] Solé, R. V. et al. (2003). Optimization in complex networks. In *Statistical mechanics of complex networks*, pages 114–126. Springer.

[Strauss, 1986] Strauss, D. (1986). On a general class of models for interaction. *SIAM review*, 28(4):513–527.

[Stumpf and Porter, 2012] Stumpf, M. P. and Porter, M. A. (2012). Critical truths about power laws. *Science*, 335(6069):665–666.

[Sweeney, 2002] Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570.

[Swendsen and Wang, 1986] Swendsen, R. H. and Wang, J.-S. (1986). Replica monte carlo simulation of spin-glasses. *Physical review letters*, 57(21):2607.

[Toda, 2012] Toda, A. A. (2012). The double power law in income distribution: Explanations and evidence. *Journal of Economic Behavior and Organization*, 84(1):364–381.

[Van Der Hofstad, 2009] Van Der Hofstad, R. (2009). *Random graphs and complex networks*, volume 1.

[van der Hoorn and Olvera-Cravioto, 2018] van der Hoorn, P. and Olvera-Cravioto, M. (2018). Typical distances in the directed configuration model. *The Annals of Applied Probability*, 28(3):1739–1792.

[Veitch and Roy, 2015] Veitch, V. and Roy, D. M. (2015). The class of random graphs arising from exchangeable random measures. *arXiv preprint arXiv:1512.03099*.

[Voitalov et al., 2019] Voitalov, I., van der Hoorn, P., van der Hofstad, R., and Krioukov, D. (2019). Scale-free networks well done. *Physical Review Research*, 1(3):033034.

[Watts and Strogatz, 1998] Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442.

[Willenborg and Waal, 2001] Willenborg, L. and Waal, T. d. (2001). Application of non-perturbative techniques for tabular data. In *Elements of Statistical Disclosure Control*, pages 175–217. Springer.

[Williamson, 2016] Williamson, S. A. (2016). Nonparametric network models for link prediction. *The Journal of Machine Learning Research*, 17(1):7102–7121.

[Willinger et al., 2009] Willinger, W., Alderson, D., and Doyle, J. C. (2009). Mathematics and the internet: A source of enormous confusion and great potential. *Notices of the American Mathematical Society*, 56(5):586–599.

[Zhang and De Sa, 2019] Zhang, R. and De Sa, C. M. (2019). Poisson-minibatching for gibbs sampling with convergence rate guarantees. *Advances in Neural Information Processing Systems*, 32.