

A Network Model characterized by a Latent Attribute Structure with Competition

Paolo Boldi*, Irene Crimaldi†, Corrado Monti‡

September 10, 2018

Abstract

The quest for a model that is able to explain, describe, analyze and simulate real-world complex networks is of uttermost practical, as well as theoretical, interest. In this paper we introduce and study a network model that is based on a latent attribute structure: each node is characterized by a number of features and the probability of the existence of an edge between two nodes depends on the features they share. Features are chosen according to a process of Indian-Buffer type but with an additional random “fitness” parameter attached to each node, that determines its ability to transmit its own features to other nodes. As a consequence, a node’s connectivity does not depend on its age alone, so also “young” nodes are able to compete and succeed in acquiring links. One of the advantages of our model for the latent bipartite “node-attribute” network is that it depends on few parameters with a straightforward interpretation. We provide some theoretical, as well experimental, results regarding the power-law behavior of the model and the estimation of the parameters. By experimental data, we also show how the proposed model for the attribute structure naturally captures most local and global properties (e.g., degree distributions, connectivity and distance distributions) real networks exhibit.

keyword: Complex network, social network, attribute matrix, Indian Buffet process

1 Introduction

Complex networks are a unifying theme that emerged in the last decades as one of the most important topics in many areas of science; the starting point is the

*Dipartimento di Informatica, Università degli Studi di Milano, Via Comelico 39/41 - 20135 Milano, Italy

†IMT Institute for Advanced Studies Lucca, Piazza San Ponziano 6, I-55100 Lucca, Italy

‡Dipartimento di Informatica, Università degli Studi di Milano, Via Comelico 39/41 - 20135 Milano, Italy

observation that many networks arising from different types of interactions (e.g., in biology, physics, chemistry, economics, technology, on-line social activity) exhibit surprising similarities that are partly still unexplained. The quest for a model that is able to explain, describe, analyze and simulate those real-world complex networks is of uttermost practical, as well as theoretical, interest.

The classical probabilistic model of graphs by Erdős and Rényi [11] soon revealed itself unfit to describe complex networks because, for example, it fails to produce a power-law degree distribution. One of the first attempts to try to obtain more realistic models was [3], where the idea of *preferential attachment* was first introduced: nodes tend to attach themselves more easily to other nodes that are already very popular, i.e. with an high number of links. Similar models were proposed by [1] and [22]. The general approach of these and other attempts is to produce probabilistic frameworks (typically with one or more parameters) giving rise to networks with statistical properties that are compatible with the ones that are observed in real-world graphs: degree distribution is just one example; other properties are degree-degree correlation, clustering coefficients, distance distribution, etc. [19].

The task of modeling the network is often undertaken directly [3, 23], but recently some authors proposed to split it into two steps (see, e.g., [24, 25]). This proposal stems from the observation that many complex networks contain two types of entities: actors on one hand, and groups (or features) on the other; every actor belongs to one, or more, groups (or can exhibit one, or more, features), and the common membership to groups (or the sharing of features) determines a relation between actors. The idea of an underlying *bipartite network* such that interpersonal connections follow from inter-group connections, derives from sociology; a seminal paper presented by Breiger [7] in 1974 described this dualism between “persons and groups”. This idea has been proved precious in social networks and their mathematical modelization [25].

In particular, many authors [26] distinguish between two kinds of models: class-based models – such as [31] – assume that every node belongs to a single class, while feature-based models use many features to describe each node. A well-known shortcoming of the first is the proliferation of classes, since dividing a class according to a new feature leads to two different classes. To overcome this limitation, classical class-based models have been extended to allow mixed membership, like in [2]. Feature-based models naturally assume this possibility. Within them, some authors (such as [18]) propose real-valued vectors to associate features to nodes; others instead assume only binary features, in which a node either exhibits a feature or it does not (see e.g. [26]). This assumption is simple and natural, and it significantly simplifies the analysis of the model.

A natural model for the evolution of such binary bipartite graphs comes from Bayesian statistics and it is known as the Indian Buffet process, introduced by Griffiths *et al.* [14, 15, 16] and, subsequently extended and studied by many authors [5, 8, 33, 34]. The process defines a plausible way for features to evolve, always according to a *rich-get-richer* principle: because of this, it represents a promising

model for affiliation networks. Since the Indian Buffet process provides a *prior distribution* in Bayesian statistics, these models have been used to reconstruct affiliation networks, with an unknown number of features, from data where only friendship relations between actors are available. An important work in this direction is [26]. However, the standard Indian Buffet process has a drawback as a model for real networks: the exchangeability assumption is often untenable in applications.

In this paper we propose and analyze a model that combines two features characterizing the evolution of a network:

1. Behind the adjacency matrix of a network there is a *latent attribute structure* of the nodes, in the sense that each node is characterized by a number of features and the probability of the existence of an edge between two nodes depends on the features they share. In other words, the adjacency matrix of a network hides a bipartite network describing the attributes of the nodes.
2. Not all nodes are equally successful in transmitting their own attributes to the new nodes. Each node n is characterized by a *random fitness parameter* R_n describing its ability to transmit the node’s attributes: the greater the value of the random variable R_n , the greater the probability that a feature of n will also be a feature of a new node, and so the greater the probability of the creation of an edge between n and the new node. Consequently, a node’s connectivity does not depend on its age alone (so also “young” nodes are able to compete and succeed in acquiring links). We refer to this aspect as *competition*.

We shape the first aspect by the definition of a model which connects the pair of attribute-vectors of two nodes, say i and j , to the probability of the existence of an edge between i and j . Other examples, which are related to the Bayesian framework based on the standard Indian Buffet model, can be found in [26, 27, 28, 30].

We model the second aspect by the definition of a stochastic dynamics for a bipartite “node-attribute” network, where the probability that a new node exhibits a certain attribute depends on the ability, represented by some random fitness parameter, of the previous nodes possessing that attribute in transmitting it. It is worthwhile to underline that in our model, as in the standard Indian Buffet process, the collection of attributes is potentially unbounded. Thus, we do not need to specify a maximum number of latent attributes *a priori*.

We were inspired by the recent generalization of the Indian buffet process presented in [5]. However, the model presented here is in some sense simpler since the parameters (that will be introduced and analyzed in the next sections) play a role that is clearer and more intuitive. Specifically, we have two parameters (α and β) that control the number of new attributes each new node exhibits (in particular $\beta > 0$ tunes the power-law behaviour of the overall number of different observed attributes), whereas the random fitness parameters R_i impact on the probability of the new nodes to inherit the attributes of the previous nodes. With respect to the model in [5], we lose some mathematical properties, but we will show that some

important results still hold true and they allow us to estimate the parameters and, in particular, the exponent of the power-law behavior.

Regarding the use of fitness parameters, we recall the work by Bianconi and Barabási [6] that introduced some fitness parameters describing the ability of the nodes to compete for links. The difference between their model and ours consists in the fact that in [6] the fitness parameters appear explicitly in the edge-probabilities; while in our model they affect the evolution of the attribute matrix and then play an implicit role in the evolution of the connections.

Summing up, the present work have different aims: firstly, we propose a simple model for the latent bipartite “node-attribute” network, where the role played by the single parameters is straightforward and easy to be interpreted; secondly, differently from other network models based on the standard Indian Buffet process, we take into account the aspect of competition and, like in [5], we introduce random fitness parameters so that nodes have a different relevance in transmitting their features to the next nodes; finally, we provide some theoretical, as well experimental, results regarding the power-law behavior of the model and the estimation of the parameters. By experimental data, we will also show how the proposed attribute structure naturally leads to a complex network model.

The paper is structured as follows. In Section 2, we introduce a model for the evolution of the attribute matrix and we provide theoretical results and tools regarding the estimation and the analysis of the quantities characterizing the model. These methods are then tested by simulations in Section 3. In order to produce a graph out of the attribute structure, in Section 4 we illustrate different models for the edge-probabilities that are based on the attribute matrix. The properties of the generated graphs are studied by simulation in Section 5. Finally, in Section 6 we analyze a real dataset and, then, in Section 7 we sum up the main novelties and merits of our work and we illustrate some possible future lines of research.

2 A model for the evolution of the attribute matrix

We assume that the nodes enter the network sequentially so that node i represents the node that comes into the network at time i . Let \mathcal{X} be an unbounded collection of possible attributes that a node can exhibit. (This means that we do not specify the total number of possible attributes *a priori*.) Each node is assumed to have only a finite number of attributes and different nodes can share one or more attributes.

Let Z be a binary bipartite network where each row Z_n represents the attributes of the node n : $Z_{n,k} = 1$ if node n has attribute k , $Z_{n,k} = 0$ otherwise. We assume that each Z_n remains unchanged in time, in the sense that every node decides its own features (attributes) when it arrives and then it will never change them thereafter. This assumption is quite natural in many contexts, e.g. in genetics.

In all the sequel we postulate that Z is left-ordered. This means that in the first

row the columns for which $Z_{1,k} = 1$ are grouped on the left and so, if the first node has N_1 features, then the columns of Z with index $k \in \{1, \dots, N_1\}$ represent these features. The second node could have some features in common with the first node (those corresponding to indices k such that $k = 1, \dots, N_1$ and $Z_{2,k} = 1$) and some, say N_2 , new features. The latter are grouped on the right of the sets for which $Z_{1,k} = 1$, i.e., the columns of Z with index $k \in \{N_1 + 1, \dots, N_2\}$ represent the new features brought by the second node. This grouping structure persists throughout the matrix Z .

Here is an example of a Z matrix with $n = 4$ nodes; in gray we show the new features adopted by each node ($N_1 = 3$, $N_2 = 2$, $N_3 = 3$, $N_4 = 2$ in this example); observe that, for every node i , the i -th row contains 1's for all the columns with indices $k \in \{N_1 + \dots + N_{i-1} + 1, \dots, N_1 + \dots + N_i\}$ (they represent the new features brought by i); moreover some elements of the columns with indices $k \in \{1, \dots, N_1 + \dots + N_{i-1}\}$ are also 1's (features brought by previous nodes and that also node i decided to adopt):

$$\begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 1 \end{pmatrix}$$

We will describe the dynamics using a culinary metaphor (similarly to what some authors do for other models, see Chinese Restaurant [29], Indian Buffet process [15, 16, 33] and their generalizations [4, 5]). We identify the nodes with the customers of a restaurant and the attributes with the dishes, so that the dishes tried by a customer represent the attributes that a node exhibits.

Fix $\alpha > 0$ and $\beta \in (-\infty, 1]$. Also, let $\text{Poi}(\lambda)$ denote the Poisson distribution with mean $\lambda > 0$. Customer (node) n is attached a random weight (that we call, in accordance with the usage in Network Theory, *fitness parameter*) R_n . We assume that each R_n is independent of R_1, \dots, R_{n-1} and of the dishes (features) experimented by customers $1, \dots, n$. The fitness parameter R_n affects the choices of the future customers (those after n), while the choices of customer n are affected by the fitness parameters and the choices of the previous ones. Indeed, it may be the case that different customers have different relevance, due to some random cause, that does not affect their choices but is relevant to the choices of future customers (i.e., their capacity of being followed).

The dynamics is as follows. Customer (node) 1 tries N_1 dishes (features), where N_1 is $\text{Poi}(\alpha)$ -distributed. For each $n \geq 1$, let S_n be the collection of dishes (features) experimented by the first n customers (nodes). For the customers following the first one, we have that:

- Customer $n + 1$ selects a subset S_n^* of S_n . Each $k \in S_n$ is included or not into

S_n^* independently of the other members of S_n . The inclusion probability is

$$P_n(k) = \frac{\sum_{i=1}^n R_i Z_{i,k}}{\sum_{i=1}^n R_i}, \quad (2.1)$$

where $Z_{i,k} = 1$ if {customer i has selected dish k } and $Z_{i,k} = 0$ otherwise. It is a preferential attachment rule: the larger the weight of a dish k at time n (given by the numerator of (2.1), i.e., the total value of the random variables R_i associated to the customers that have chosen it until time n), the greater the probability that it will be chosen by the future customer $n + 1$.¹

- In addition to S_n^* , customer $n + 1$ also tries N_{n+1} new dishes, where N_{n+1} is Poi(Λ_n)-distributed with

$$\Lambda_n = \frac{\alpha}{(\sum_{i=1}^n R_i)^{1-\beta}}. \quad (2.2)$$

For each k in S_{n+1} , the matrix element $Z_{n+1,k}$ is set equal to 1 if customer $n + 1$ has selected dish k , equal to zero otherwise.

Besides the assumption of independence, we also assume that the random parameters R_n are identically distributed with $R_n \geq v$ for each n and a certain number $v > 0$, and $E[R_n^2] < +\infty$.

We set $E[R_n] = m_R$ and $L_n = \text{card}(S_n) = \sum_{i=1}^n N_i$, i.e.

L_n = overall number of different dishes experimented by the first n customers
= overall number of different observed attributes for the first n nodes.

In the previous example, we have $L_1 = 3, L_2 = 5, L_3 = 8, L_4 = 10$.

The meaning of the parameters is the following. The random fitness parameters R_n fundamentally control the probability of transmitting the attributes to the new nodes. The main effect of β is that it regulates the asymptotic behavior of the random variable L_n (see Theorem 2.1). In particular, $\beta > 0$ is the power-law exponent of L_n . The main effect of α is the following: the larger α , the larger the total number of new tried dishes by a customer (and so the larger the total number of 1's in a row of the binary matrix Z). It is worth to note that β fits the asymptotic behaviour of L_n (in particular, the power-law exponent of L_n) and, separately, α fits the number of observed features.

¹As we will discuss in the conclusions (Sec. 7), we can generalize our model by introducing another parameter $c \geq 0$ in the inclusion probabilities so that

$$P_n(k) = \frac{\sum_{i=1}^n R_i Z_{i,k}}{c + \sum_{i=1}^n R_i}.$$

For the moment, we set $c = 0$. Note that this choice implies $P_n(k) = 1$ for all n and $k = 1, \dots, N_1$. Therefore, we could consider the first node and its features “fictitious”, in the sense that the “true” dynamics is for $n \geq 2$ and $k \geq N_1 + 1$.

The mathematical formalization of the above model can be performed by means of random measures [21] with atoms corresponding to the tried dishes (observed attributes), similarly to [5, 8, 34]. More precisely, besides the sequence of positive real random variables (R_n) , we can define a sequence of random measures (M_n) , such that each M_{n+1} is, conditionally on the past $(M_i, R_i : i \leq n)$, a Bernoulli random measure with a hazard measure ν_n , having a discrete part related to the points k in S_n and their weights $P_n(k)$ and a diffuse part with total mass equal to Λ_n .

2.1 Theoretical results regarding the estimation of the parameters α and β

In this section we prove some properties regarding the asymptotic behavior of L_n . In particular, the first result shows a logarithmic behavior for $\beta = 0$ and a power-law behavior for $\beta \in (0, 1]$. These results allow us to define suitable estimators for β and α .

Theorem 2.1. *Using the previous notation, the following statements hold true:*

- a) $\sup_n L_n = L < +\infty$ a.s. for $\beta < 0$;
- b) $L_n/\ln(n) \xrightarrow{a.s.} \alpha/m_R$ for $\beta = 0$;
- c) $L_n/n^\beta \xrightarrow{a.s.} \alpha/(\beta m_R^{1-\beta})$ for $\beta \in (0, 1]$.

Proof. Let us prove assertion a), first. Let \mathcal{F}_i be the natural σ -field associated to the model until time i and set $\Lambda_0 = \alpha$. Since, conditionally on \mathcal{F}_i , the distribution of N_{i+1} is $\text{Poi}(\Lambda_i)$, we have

$$P(N_{i+1} \geq 1) = E[P(N_{i+1} \geq 1 \mid \mathcal{F}_i)] \leq E[\Lambda_i].$$

Since $R_i \geq v > 0$, we obtain

$$\sum_i P(N_{i+1} \geq 1) \leq \alpha \sum_i \frac{1}{(vi)^{(1-\beta)}} < +\infty$$

(where the convergence of the series is due to the assumption $\beta < 0$). By the Borel-Cantelli lemma, we conclude that

$$P(N_i > 0 \text{ infinitely often}) = P(N_i \geq 1 \text{ infinitely often}) = 0.$$

Hence, if $\beta < 0$, there is a random index N such that $L_n = L_N$ a.s. for all $n \geq N$, which concludes the proof of a).

The assertion c) is trivial for $\beta = 1$ since, in this case, L_n is the sum of n independent random variables with distribution $\mathcal{P}(\alpha)$ and so, by the classical strong law of large numbers, $L_n/n \xrightarrow{a.s.} \alpha$.

Now, let us prove assertions b) and c) for $\beta \in [0, 1)$. Define

$$\lambda(\beta) = \frac{\alpha}{m_R} \text{ if } \beta = 0 \quad \text{and} \quad \lambda(\beta) = \frac{\alpha}{\beta m_R^{1-\beta}} \text{ if } \beta \in (0, 1),$$

$$a_n(\beta) = \log n \text{ if } \beta = 0 \quad \text{and} \quad a_n(\beta) = n^\beta \text{ if } \beta \in (0, 1).$$

We need to prove that

$$\frac{L_n}{a_n(\beta)} \xrightarrow{\text{a.s.}} \lambda(\beta).$$

First, we observe that we can write

$$\frac{\sum_{i=1}^{n-1} \Lambda_i}{a_n(\beta)} = \alpha \frac{\sum_{i=1}^{n-1} i^{\beta-1} (\bar{R}_i)^{\beta-1}}{a_n(\beta)},$$

where, by the strong law of the large numbers,

$$\bar{R}_i = \frac{\sum_{j=1}^i R_j}{i} \xrightarrow{\text{a.s.}} m_R.$$

Therefore, since $\sum_{i=1}^{n-1} i^{\beta-1}/a_n(\beta)$ converges to 1 when $\beta = 0$ and to $1/\beta$ when $\beta \in (0, 1)$, we get

$$\frac{\sum_{i=1}^{n-1} \Lambda_i}{a_n(\beta)} \xrightarrow{\text{a.s.}} \lambda(\beta). \quad (2.3)$$

Next, let us define

$$T_0 = 0 \quad \text{and} \quad T_n = \sum_{i=1}^n \frac{N_i - E[N_i | \mathcal{F}_{i-1}]}{a_i(\beta)} = \sum_{i=1}^n \frac{N_i - \Lambda_{i-1}}{a_i(\beta)}.$$

Then, (T_n) is a martingale with respect to (\mathcal{F}_n) and

$$E[T_n^2] = \sum_{i=1}^n \frac{E[(N_i - \Lambda_{i-1})^2]}{a_i(\beta)^2} = \sum_{i=1}^n \frac{E\{E[(N_i - \Lambda_{i-1})^2 | \mathcal{F}_{i-1}]\}}{a_i(\beta)^2} = \sum_{i=1}^n \frac{E[\Lambda_{i-1}]}{a_i(\beta)^2}.$$

Since $R_i \geq v > 0$, it is easy to verify that $E[\Lambda_i] = O(i^{-(1-\beta)})$ and so $\sup_n E[T_n^2] = \sum_{i=1}^\infty \frac{E[\Lambda_{i-1}]}{a_i(\beta)^2} < \infty$. Thus, (T_n) converges a.s., and the Kronecker's lemma implies

$$\frac{1}{a_n(\beta)} \sum_{i=1}^n a_i(\beta) \frac{(N_i - \Lambda_{i-1})}{a_i(\beta)} \xrightarrow{\text{a.s.}} 0,$$

so finally

$$\lim_n \frac{L_n}{a_n(\beta)} = \lim_n \frac{\sum_{i=1}^n N_i}{a_n(\beta)} = \lim_n \frac{\sum_{i=1}^n \Lambda_{i-1}}{a_n(\beta)} = \lim_n \frac{\Lambda_0 + \sum_{i=1}^{n-1} \Lambda_i}{a_n(\beta)} = \lambda(\beta) \quad \text{a.s.} \quad (2.4)$$

■

The above result entails that $\ln(L_n)/\ln(n)$ is a strongly consistent estimator of $\beta \in [0, 1]$. In fact:

- if $\beta = 0$ then $L_n \stackrel{a.s.}{\sim} \frac{\alpha}{m_R} \ln(n)$ as $n \rightarrow +\infty$; hence $\ln(L_n) \stackrel{a.s.}{\sim} \ln(\alpha/m_R) + \ln(\ln(n))$, therefore $\ln(L_n)/\ln(n) \stackrel{a.s.}{\sim} \ln(\alpha/m_R)/\ln(n) + \ln(\ln(n))/\ln(n) \xrightarrow{a.s.} 0 = \beta$;
- if $\beta > 0$, we have $L_n \stackrel{a.s.}{\sim} \lambda(\beta)n^\beta$ as $n \rightarrow +\infty$ so $\ln(L_n) \stackrel{a.s.}{\sim} \ln(\lambda(\beta)) + \beta \ln(n)$, hence $\ln(L_n)/\ln(n) \stackrel{a.s.}{\sim} \ln(\lambda(\beta))/\ln(n) + \beta \xrightarrow{a.s.} \beta$.

Remark 2.2. In practice, the value of $\ln(L_n)/\ln(n)$ may be quite far from the limit value β when n is small. Hence, it may be worth trying to fit the power-law dependence of L_n as a function of n with standard techniques [10] and use the slope $\widehat{\beta}_n$ of the regression line in the log-log plot as an effective estimator for β .

Finally, assuming that $\beta \in [0, 1]$ and m_R are known, we can get a strongly consistent estimator of α , as:

$$m_R \frac{L_n}{\ln(n)} \quad \text{for } \beta = 0 \quad \text{and} \quad m_R^{1-\beta} \beta \frac{L_n}{n^\beta} \quad \text{for } 0 < \beta \leq 1.$$

In practice, we assume β equal to the estimated value $\widehat{\beta}_n$ (as defined before) and we take m_R equal to the estimated value $\bar{R}_n = \sum_{i=1}^n R_i/n$, if the random parameters R_i are known. In Section 3.2, we will discuss the case when the random variables R_i are unknown.

Remark 2.3. Once more, it may be better in practice to estimate α as

$$\begin{aligned} \widehat{\alpha}_n &= m_R \widehat{\gamma}_n & \text{when } \beta = 0 \\ \widehat{\alpha}_n &= \beta m_R^{1-\beta} \widehat{\gamma}_n & \text{when } 0 < \beta \leq 1, \end{aligned} \tag{2.5}$$

where $\widehat{\gamma}_n$ is the slope of the regression line in the plot $(\ln(n), L_n)$ or in the plot (n^β, L_n) according to whether $\beta = 0$ or $\beta \in (0, 1]$.

We complete this section with a central-limit theorem that gives the rate of convergence of $L_n/a_n(\beta)$ to $\lambda(\beta)$ when $\beta \in [0, 1]$.

Theorem 2.4. *If $\beta \in [0, 1]$, then we have the following convergence in distribution²:*

$$\sqrt{a_n(\beta)} \left\{ \frac{L_n}{a_n(\beta)} - \lambda(\beta) \right\} \xrightarrow{d} \mathcal{N}(0, \lambda(\beta)).$$

Proof. The result for $\beta = 1$ follows from the classical central limit theorem, since, in this case, L_n is the sum of n independent random variables with distribution $\mathcal{P}(\alpha)$. Assume now $\beta \in [0, 1)$ and set $\Lambda_0 = \alpha$. We first prove that

$$\sqrt{a_n(\beta)} \left\{ \frac{\sum_{i=1}^n \Lambda_{i-1}}{a_n(\beta)} - \lambda(\beta) \right\} \xrightarrow{P} 0. \tag{2.6}$$

²Actually, the convergence is in the sense of the *stable* convergence, which is stronger than the convergence in distribution. Indeed, stable convergence is a form of convergence intermediate between convergence in distribution and convergence in probability.

By some calculations, condition (2.6) is equivalent to

$$\frac{\sum_{i=1}^{n-1} \left\{ \left(\sum_{j=1}^i R_j \right)^{\beta-1} - (m_R i)^{\beta-1} \right\}}{\sqrt{a_n(\beta)}} \xrightarrow{P} 0. \quad (2.7)$$

Since $R_j \geq v > 0$, we have $m_R \geq v > 0$ and we obtain

$$\begin{aligned} E \left[\left| (m_R i)^{\beta-1} - \left(\sum_{j=1}^i R_j \right)^{\beta-1} \right| \right] &\leq \frac{E \left[\left| \left(\sum_{j=1}^i R_j \right)^{1-\beta} - (m_R i)^{1-\beta} \right| \right]}{(v i)^{2(1-\beta)}} \\ &\leq \frac{1}{(v i)^{2(1-\beta)}} \frac{1-\beta}{(v i)^\beta} E \left[\left| \sum_{j=1}^i R_j - m_R i \right| \right] \\ &= \frac{1-\beta}{v^{2-\beta}} \frac{1}{i^{1-\beta}} E [|\bar{R}_i - m_R|] \\ &\leq \frac{1-\beta}{v^{2-\beta}} \frac{1}{i^{1-\beta}} \sqrt{\text{Var}[\bar{R}_i]} = \frac{(1-\beta) \sqrt{\text{Var}[R_1]}}{v^{2-\beta}} \frac{i^{\beta-1}}{\sqrt{i}}. \end{aligned}$$

This proves condition (2.7) (and so (2.6)). Indeed, we have

$$\begin{aligned} &\frac{1}{\sqrt{a_n(\beta)}} E \left[\left| \sum_{i=1}^{n-1} \left\{ \left(\sum_{j=1}^i R_j \right)^{\beta-1} - (m_R i)^{\beta-1} \right\} \right| \right] \leq \\ &\frac{1}{\sqrt{a_n(\beta)}} \sum_{i=1}^{n-1} E \left[\left| (m_R i)^{\beta-1} - \left(\sum_{j=1}^i R_j \right)^{\beta-1} \right| \right] \leq \\ &\frac{(1-\beta) \sqrt{\text{Var}[R_1]}}{v^{2-\beta}} \frac{1}{\sqrt{a_n(\beta)}} \sum_{i=1}^{n-1} \frac{1}{i^{1-(\beta-1/2)}} \rightarrow 0. \end{aligned}$$

Next, define

$$T_n = \sqrt{a_n(\beta)} \left\{ \frac{L_n}{a_n(\beta)} - \frac{\sum_{i=1}^n \Lambda_{i-1}}{a_n(\beta)} \right\} = \frac{\sum_{i=1}^n (N_i - \Lambda_{i-1})}{\sqrt{a_n(\beta)}}.$$

In view of (2.6), it suffices to show that $T_n \xrightarrow{d} \mathcal{N}(0, \lambda(\beta))$.

To this end, for $n \geq 1$ and $i = 1, \dots, n$, define

$$T_{n,i} = \frac{N_i - \Lambda_{i-1}}{\sqrt{a_n(\beta)}}, \quad \mathcal{G}_{n,0} = \mathcal{F}_0 \quad \text{and} \quad \mathcal{G}_{n,i} = \mathcal{F}_i,$$

where \mathcal{F}_i is the natural σ -field associated to the model until time i . Then, we have $E[T_{n,i} | \mathcal{G}_{n,i-1}] = 0$, $\mathcal{G}_{n,i} \subseteq \mathcal{G}_{n+1,i}$ and $T_n = \sum_{i=1}^n T_{n,i}$. Thus, by a martingale central limit theorem (see [17]), $T_n \xrightarrow{d} \mathcal{N}(0, \lambda(\beta))$ provided

$$(i) \sum_{i=1}^n T_{n,i}^2 \xrightarrow{P} \lambda(\beta), \quad (ii) \max_{1 \leq i \leq n} |T_{n,i}| \xrightarrow{P} 0, \quad (iii) \sup_n E \left[\max_{1 \leq i \leq n} T_{n,i}^2 \right] < \infty.$$

Let

$$D_i = (N_i - \Lambda_{i-1})^2 \quad \text{and} \quad U_n = \frac{\sum_{i=1}^n \{D_i - E[D_i | \mathcal{F}_{i-1}]\}}{a_n(\beta)} = \frac{\sum_{i=1}^n (D_i - \Lambda_{i-1})}{a_n(\beta)}.$$

By the same martingale argument used in the proof of the previous theorem and by Kronecker's lemma, $U_n \xrightarrow{a.s.} 0$. Then, by (2.3),

$$\sum_{i=1}^n T_{n,i}^2 = \frac{\sum_{i=1}^n D_i}{a_n(\beta)} = U_n + \frac{\sum_{i=1}^n \Lambda_{i-1}}{a_n(\beta)} \xrightarrow{a.s.} \lambda(\beta).$$

This proves condition (i). As to (ii), fix $k \geq 1$ and note that

$$\max_{1 \leq i \leq n} T_{n,i}^2 \leq \frac{\max_{1 \leq i \leq k} D_i}{a_n(\beta)} + \max_{k < i \leq n} \frac{D_i}{a_i(\beta)} \leq \frac{\max_{1 \leq i \leq k} D_i}{a_n(\beta)} + \sup_{i > k} \frac{D_i}{a_i(\beta)} \quad \text{for } n > k.$$

Hence, $\limsup_n \max_{1 \leq i \leq n} T_{n,i}^2 \leq \limsup_n \frac{D_n}{a_n(\beta)}$ and condition (ii) follows since

$$\frac{D_n}{a_n(\beta)} = \frac{\sum_{i=1}^n D_i}{a_n(\beta)} - \frac{\sum_{i=1}^{n-1} D_i}{a_n(\beta)} \xrightarrow{a.s.} 0.$$

Finally, condition (iii) is a consequence of

$$\begin{aligned} E \left[\max_{1 \leq i \leq n} T_{n,i}^2 \right] &\leq \frac{\sum_{i=1}^n E[D_i]}{a_n(\beta)} = \frac{\sum_{i=1}^n E[\Lambda_{i-1}]}{a_n(\beta)} = \\ &= \frac{\Lambda_0 + \sum_{i=1}^{n-1} E[\Lambda_i]}{a_n(\beta)} \leq \frac{\alpha (1 + \sum_{i=1}^{n-1} (vi)^{\beta-1})}{a_n(\beta)}. \end{aligned}$$

■

2.2 Analysis of the random fitness parameters R_i

Now our purpose is to find, under the assumption of our model, a procedure to get information on the random variables R_i from the data, that typically are the values of Z_1, \dots, Z_n , i.e., n rows of the matrix Z , where n is the number of the observed nodes.

Unfortunately, this goal is not easily tractable as we will point out in the sequel. The method we empirically tested extracts from the data, with a maximum log-likelihood procedure (see Section 3.2), a plausible realization $\hat{r}_1, \dots, \hat{r}_{k_n}$ of R_1, \dots, R_{k_n} , for a suitable k_n ; this information could be useful, for instance, to reconstruct the ranking induced by R_i . Note that we ideally would like to find a probable realization for all the fitness parameters of the observed nodes (not only for the first k_n nodes), but we do not possess the same amount of information about all R_i : in particular, while R_1 influences all the subsequent observed rows of the

matrix Z , R_{n-1} has only influence over Z_n . So we cannot expect to find good values for all the random variables.

With the above purpose in mind, we now give a general expression for the conditional probability of observing $Z_1 = z_1, \dots, Z_n = z_n$ given R_1, \dots, R_{n-1} . We refer to Section 2 for the notation.

The first row Z_1 is simply identified by $L_1 = N_1$ and so

$$\begin{aligned} P(Z_1 = z_1) &= P(N_1 = n_1 = \text{card}\{k : z_{1,k} = 1\}) \\ &= \text{Poi}(\alpha)\{n_1\} = e^{-\alpha} \frac{\alpha^{n_1}}{n_1!}. \end{aligned}$$

Then the second row is identified by the values $Z_{2,k}$ with $k = 1, \dots, L_1 = N_1$ and by N_2 and so

$$\begin{aligned} P(Z_2 = z_2 | Z_1, R_1) &= \\ P(Z_{2,k} = z_{2,k} \text{ for } k = 1, \dots, L_1, N_2 = n_2 = \text{card}\{k > L_1 : z_{2,k} = 1\} | Z_1, R_1) &= \\ \prod_{k=1}^{L_1} P_1(k)^{z_{2,k}} (1 - P_1(k))^{1-z_{2,k}} \times \text{Poi}(\Lambda_1)\{n_2\}, \end{aligned}$$

where $P_1(k)$ is defined in (2.1) and Λ_1 is defined in (2.2).

The general formula is

$$\begin{aligned} P(Z_{j+1} = z_{j+1} | Z_1, R_1, \dots, Z_j, R_j) &= \\ P(Z_{j+1,k} = z_{j+1,k} \text{ for } k = 1, \dots, L_j, \\ N_{j+1} = n_{j+1} = \text{card}\{k > L_j : z_{j+1,k} = 1\} | Z_1, R_1, \dots, Z_j, R_j) &= \\ \prod_{k=1}^{L_j} P_j(k)^{z_{j+1,k}} (1 - P_j(k))^{1-z_{j+1,k}} \times \text{Poi}(\Lambda_j)\{n_{j+1}\}, \end{aligned}$$

where $P_j(k)$ is defined in (2.1) and Λ_j is defined in (2.2).

Hence, for n nodes, we can write a formula for the conditional probability of observing $Z_1 = z_1, \dots, Z_n = z_n$ given R_1, \dots, R_{n-1} :

$$\begin{aligned} P(Z_1 = z_1, \dots, Z_n = z_n | R_1, \dots, R_{n-1}) &= \\ P(Z_1 = z_1) \prod_{j=1}^{n-1} P(Z_{j+1} = z_{j+1} | Z_1, R_1, \dots, Z_j, R_j). \end{aligned} \tag{2.8}$$

2.2.1 A Monte Carlo method

The algorithm we applied is essentially a MCMC (Markov Chain Monte Carlo) method [13], which uses the basic principle of Gibbs sampling [9]: fix all components of a vector except one and compare the different values of the likelihood obtained for various values of the non-fixed component.

The method employs the aforementioned formula (2.8) for the conditional probability of observing $Z_1 = z_1, \dots, Z_n = z_n$ given the values of R_1, \dots, R_{n-1} . Precisely, using the symbol \bar{z} in order to denote the matrix with rows z_1, \dots, z_n and the symbol \bar{r} in order to denote a vector of component r_1, \dots, r_n , set

$$P(Z = \bar{z} | R = \bar{r}) = P(Z_1 = z_1, \dots, Z_n = z_n | R_1 = r_1, \dots, R_n = r_n). \quad (2.9)$$

We want to find a vector \hat{r} that is a point maximizing the likelihood function (2.9) corresponding to the observed \bar{z} .³

The basic algorithm is described in Alg. 1. It is regulated by these parameters:

- $\bar{r}^0 \in \mathbb{R}^n$ is the initial guess for \hat{r} ;
- $J \in \mathbb{N}^+$ is the number of “jumps to a new value”, i.e., the number of the new values analyzed for a certain component at each step;
- $\sigma \in \mathbb{R}^+$ is the standard deviation of each “jump”.

Algorithm 1 Basic Monte Carlo algorithm to find \hat{r} .

INPUT: z_1, \dots, z_n , the observed features of each of the n observed nodes, i.e., the first n rows of the attribute matrix Z

OUTPUT: \hat{r} , a maximum point for the likelihood function associated to the input data

DESCRIPTION:

1. $\hat{r} \leftarrow \bar{r}^0$
2. Repeat the following loop until convergence:
 - (a) Choose a random node $i \in \{1, \dots, n\}$
 - (b) Extract J values h_1, \dots, h_J from the normal distribution $\mathcal{N}(r_i, \sigma^2)$; re-sample each h_j until $h_j > 0$.
 - (c) For each value h_j , compute

$$\mathcal{L}(h_j) = P(Z = \bar{z} | R_1 = r_1, \dots, R_i = h_j, \dots, R_n = r_n)$$

- (d) $\hat{r}_i \leftarrow \arg \max_{h \in \{r_i, h_1, \dots, h_J\}} \mathcal{L}(h)$

³We point out that our algorithm can not be considered a proper statistical estimation procedure for the fitness parameters. In particular, although it resembles the Bayesian *Maximum a posteriori probability* (MAP) estimation when the a priori distribution is an (improper) uniform distribution, we do not have a vector of parameters with a fixed dimension: the number of parameters in our case increases with the number of observations.

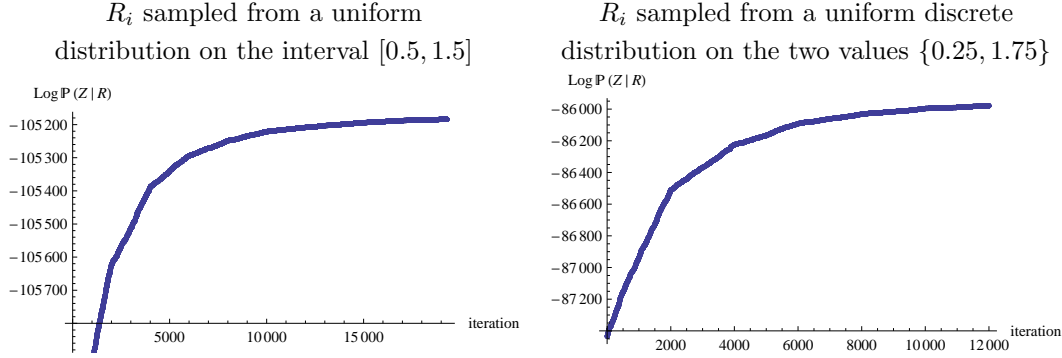


Figure 1: Value of the log-likelihood during the execution of the algorithm, for different distributions of R_i . The chosen algorithm parameters are $\sigma^2 = 1$, $J = 4$ and $\bar{r}^0 = \mathbf{1}$ (the vector with all 1's). The matrix Z has 2000 rows (nodes) and it was generated with $\alpha = 3$ and $\beta = 0.9$.

It is worth to note that, given $\mathcal{L}(r_i)$, it is possible to find $\mathcal{L}(h)$ without re-doing the whole computation. In fact, let us consider the product in eq. (2.8): a change from r_i to h must be taken into account only from the i -th factor onward – that is, for the factors that come after $P(Z_i = z_i | Z_1, R_1, \dots, Z_{i-1}, R_{i-1})$. In particular, let $\delta = h - r_i$; then, for each j -th factor, with $j \geq i$, we have to:

- add δ to the term $\sum_{i=1}^j R_i$, inside Λ_j and $P_j(k)$ (defined in eq. (2.1) and (2.2));
- add δ to the numerator of $P_j(k)$ when k is s.t. $z_{i,k} = 1$; that is to say, change the global weight of a feature only if the node we changed has that feature.

Every other term in the equation remains unchanged and does not need to be computed again. This remark allows us to speed up the implementation considerably.

Figure 1 confirms that the algorithm moves toward a vector \hat{r} maximizing $P(Z = \bar{z} | R = \bar{r})$ and shows that the algorithm effectively converges. As a stopping criterion, we can use the maximum increase in the log-likelihood in the last iterations: when this is under a certain threshold t , we stop the algorithm. The obtained outputs will be discussed in details in Section 3.2.

As already said, one point that we need to keep in mind is that we do not possess the same amount of information about all the random variables R_i : in particular, while R_1 influences all the subsequent rows of the matrix Z , R_{n-1} has only influence over the last one. So we cannot expect the output values to be very accurate for the last segment. For this reason, we also implemented a variant of the algorithm that considers only the first k_n nodes. Thus, we have another algorithm parameter k_n so that the choice of the jumping node at step 2(a) is restricted to $i \in \{1, \dots, k_n\}$

and, finally, the output will be the corresponding segment of $\widehat{\bar{r}}$, i.e., $\widehat{r}_1, \dots, \widehat{r}_{k_n}$. This variant converges faster and moreover it allows to use larger values of the algorithm parameter J .

Another relevant point is that the parameters α and β enter the expression (2.8). Therefore, in practice, before applying the algorithm, we need to estimate them. As shown in Remark 2.2, we are able to estimate β starting from the observed values of the matrix Z . On the other hand, as shown in Remark 2.3, the estimation of α presupposes the knowledge of the mean value m_R of the fitness parameters R_i (except for the special case $\beta = 1$). Hence, we are in the situation in which, in order to get information on the fitness parameters by the proposed algorithm, we need to estimate α and β , but, in order to estimate α , we need to know the mean value m_R . This problem can be partially solved as follows. Since the term $P(Z_1 = z_1)$ does not contain the R_i 's, the research of a vector $\widehat{\bar{r}}$ that maximizes (2.9) is equivalent to the research of a vector $\widehat{\bar{r}}$ maximizing the product

$$\prod_{j=1}^{n-1} P(Z_{j+1} = z_{j+1} | Z_1, R_1, \dots, Z_j, R_j)$$

in formula (2.8). On the other hand, each term of the above product contains the inclusion probabilities $P_j(k)$, that are invariant with respect to the normalization of the R_i 's by their mean value m_R , and the Λ_j 's that have the property

$$\Lambda_j = f(\alpha, \beta, \bar{r}) = f(\alpha/(m_R)^{1-\beta}, \beta, \bar{r}/m_R)$$

(where \bar{r}/m_R denotes the vector with components r_i/m_R). Consequently, starting from the observed values of the matrix Z , we can

- first, estimate β by Remark 2.2;
- then estimate $\alpha' = \alpha/(m_R)^{1-\beta}$ by Remark 2.3 (i.e., $\widehat{\alpha}'_n$ equal to $\widehat{\gamma}_n$ or $\beta \widehat{\gamma}_n$ according to the estimated value of β);
- finally, extract a plausible realization $\widehat{\bar{r}}' = \widehat{\bar{r}}/m_R$ (of the random variables $R'_i = R_i/m_R$) as a maximum point of the corresponding expression of the likelihood with the estimated value of β and α' .

Therefore, the output of the algorithm will be α' , β and a plausible realization $\widehat{\bar{r}}'$ of the random variables $R'_i = R_i/m_R$.

Finally, we highlight that it is possible to experiment other variants of the algorithm, for example, by using a distribution different from the normal for the jumps, or changing σ during the execution (e.g., reducing it according to some “cooling schedule”, as it happens in simulated annealing [12]). Additionally, instead of looking for the values on the whole positive real line, we could restrict the research on a suitable interval (guessed for the particular real case).

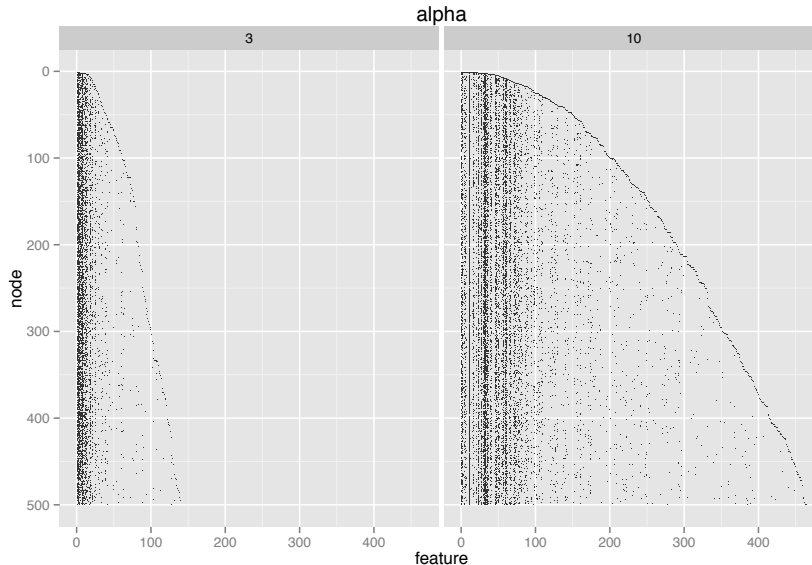


Figure 2: The Z matrix for $n = 500$, two different values of α ($\alpha = 3$ and $\alpha = 10$) and a fixed $\beta = 0.5$. The random variables R_i are uniformly distributed on the interval $[0.25, 1.75]$.

3 Simulations for the attribute matrix

In this section, we shall present a number of simulations we performed in order to illustrate the role of the parameters of the model and also to see how good the proposed tools turn out to be.

3.1 Estimating α and β

Firstly, we aim at pointing out the role played by the model parameters α and β . Therefore, we fix a distribution for the random fitness parameters with $m_R = 1$ and we simulate the matrix Z for different values of α and β (fixing one and making the other one change). More precisely, we assume that the random variables R_n are uniformly distributed on the interval $[0.25, 1.75]$.

In Figure 2, we visualize the effect of α : a larger α yields a larger number of new attributes per node.

In Figure 3, instead, we visualize how different positive values of β yield a different power-law (asymptotic) behavior of L_n . Indeed, in this figure, we have the log-log plot of L_n as a function of n . In the first two panels, we present two different positive values of β (0.75 and 0.5), showing the correspondence with the power-law exponent of L_n , estimated by the slope of the regression line. Moreover, in the third panel, we point out that the parameter α do not affect the power-law exponent of L_n .

Figure 4 underlines that the estimator proposed in remark 2.2 works better (i.e. with a more precision) for large values of β since L_n reaches the power-law behavior more quickly for larger values of β .

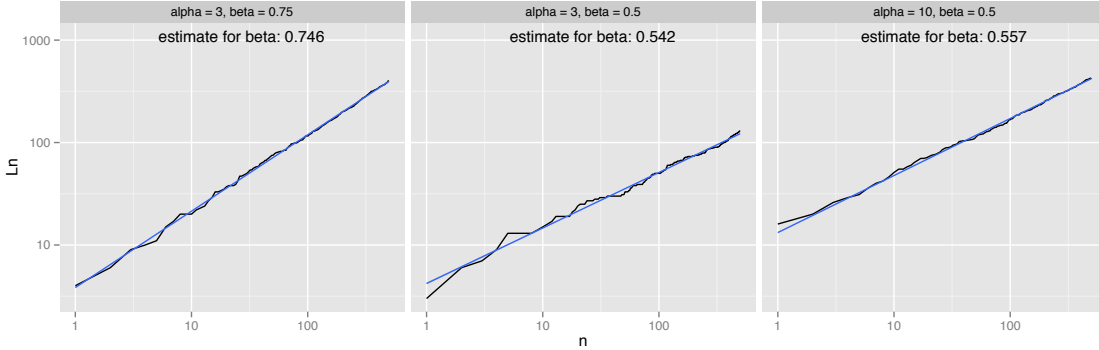


Figure 3: Correspondence between the parameter β and the power-law exponent of L_n as a function of n . The estimate of β is the slope of the regression line. Here, we have 500 nodes and the random variables R_i are uniformly distributed on the interval $[0.25, 1.75]$. Values for α and β are indicated above; we can see how different values for α do not affect the power-law behaviour.

Similarly, we evaluated the estimator $\hat{\alpha}_n$ of α , obtained by using the slope of the regression line in the plot of L_n as a function of n^β , as said in Remark 2.3 (note that we have $m_R = 1$ and so α coincides with α'). Results are illustrated in Figure 5 and show how this estimator yields good results.

We also checked how the shape of the matrix Z is influenced by the distribution of the random parameters R_n . More precisely, we analyzed the effect of ε on the shape of Z when the random variables R_i are uniformly distributed on the interval $[\varepsilon, 2 - \varepsilon]$, with $0 < \varepsilon \leq 1$, so that $E[R_i] = m_R = 1$ and the variance of R_i is $Var[R_i] = (1 - \varepsilon)^2/3$, which goes to zero as $\varepsilon \rightarrow 1$. Hence, when ε is smaller, the variance of the R_n 's is larger, so that also a “young” nodes i have some chance of transmitting their attributes to the other nodes (recall that a larger R_i makes i more successful in transmitting its own attributes). This is witnessed (see Figure 6) by the number of “blackish” vertical lines, that are more or less widespread in the whole spectrum of nodes; whereas for larger ε they are more concentrated on the left-hand side (i.e., only the first nodes successfully transmit their attributes).

3.2 Analysis of the random fitness parameters R_i

We proceeded to test empirically how the Monte Carlo method performs in recovering the information on the fitness parameters R_i . We tested its behavior against various distributions of R_i ; specifically, a uniform distribution on an interval, a two-class uniform distribution, and finally a discrete power-law distribution with 10

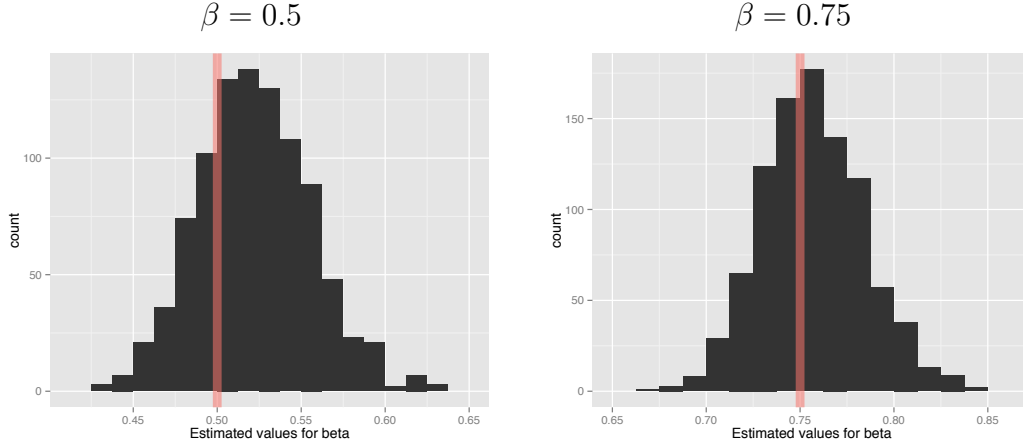


Figure 4: Distribution of the estimator $\hat{\beta}_n$ of β over 1000 experiments, each with $n = 2000$ and $\alpha = 3$. The random variables R_i are uniformly distributed on the interval $[0.25, 1.75]$. The red line indicates the true value of β .

classes. In the following of this section we illustrate the details of such experiments, while, in the next section, we will try to measure the performance of the proposed technique.

In every experiment, the matrix Z has $n = 2000$ nodes and it was generated with $\alpha = 3$ and $\beta = 0.9$. The Monte Carlo algorithm parameters were set as follows: $\sigma^2 = 1$, $J = 4$ and $\vec{r}^0 = \mathbf{1}$ (the vector with all 1's).

For the first experiment, each R_i is sampled from the uniform distribution on the interval $[0.5, 1.5]$. We used the previously discussed techniques to find the estimates of α and β : the estimated values are $\hat{\alpha} = 3.095$ and $\hat{\beta} = 0.893$ (note that we have $m_R = 1$ and so $\alpha = \alpha'$ and $\hat{r} = \hat{r}'$). Then, we tried the proposed Monte Carlo algorithm with the stopping threshold $t = 1/4$. Results are visualized in Figure 7, according to two different orderings of the nodes:

- i) in the natural order, so that we confirm that our predictions are better for the first (i.e., the oldest) nodes than for the last (i.e., the youngest) ones;
- ii) ordered by their true fitness values, so that we can show that we are, more or less, able to reconstruct the relative order of the fitness parameters (this fact will be made clearer in Section 3.3).

In the second experiment, we applied our algorithm to a discrete case: we sampled the fitness parameters R_i from a set of only two values, $\{0.25, 1.75\}$, each with probability $\frac{1}{2}$. We left the parameters of the model and the ones of the algorithm unaltered, except for moving the stopping threshold t from $1/4$ to 1 . The estimated

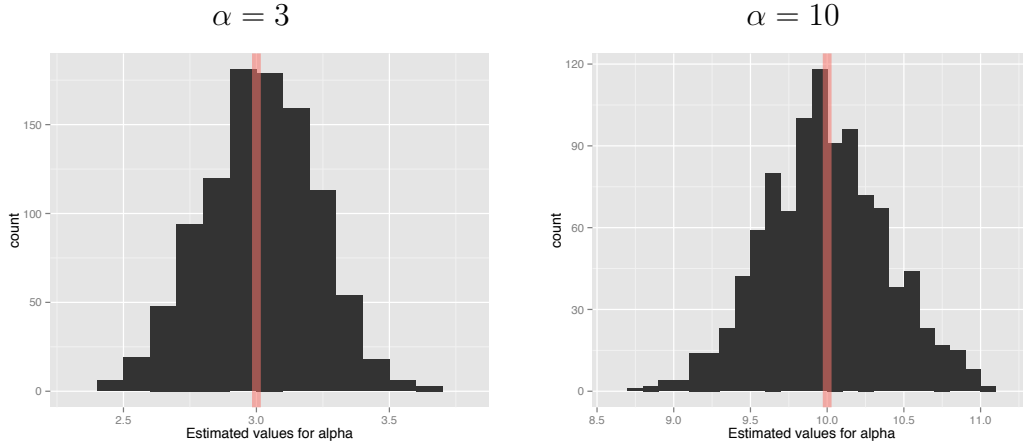


Figure 5: Distribution of the estimator $\hat{\alpha}_n$ of α over 1000 experiments, each with $n = 2000$ and $\beta = 0.5$. The random variables R_i are uniformly distributed on the interval $[0.25, 1.75]$ (so $m_R = 1$ and $\alpha = \alpha'$). The red line indicates the true value of α .

values for $\alpha = 3$ and $\beta = 0.9$ are, respectively, $\hat{\alpha} = 2.922$ (again $m_R = 1$ and so $\alpha = \alpha'$ and $\hat{r} = \tilde{r}'$) and $\hat{\beta} = 0.903$. The results of this second experiment are more encouraging (we will see precise measurements in Section 3.3). In this case, the output values of the algorithm are closer to the true ones (see Figure 8). Moreover, we can still observe the same phenomena, i) and ii), described above.

Finally, we applied the algorithm to a third case: we sampled R_i from a normalized power-law discrete distribution, with 10 possible values – specifically, a normalized discrete Zipf’s law with exponent 2 and number of values 10. We left both algorithm and model parameters unaltered and we used 1 as the stopping threshold t .

The estimated values for $\alpha = 3$ and $\beta = 0.9$ are, respectively, $\hat{\alpha} = 3.595$ (again $m_R = 1$ and so $\alpha = \alpha'$ and $\hat{r} = \tilde{r}'$) and $\hat{\beta} = 0.868$.

Results for this case show that – despite the fact that we have now a discrete distribution with more than two values – our approach can recover information (especially for larger fitness values), as can be seen in Figure 9 and in Section 3.3.

We conclude this section noting that, for each of the experiments, the Monte Carlo algorithm looks for the values of the fitness parameters on the whole positive real line. We would obtain better outputs if we could restrict the research on a suitable interval for each case, assuming a partial knowledge of the shape of the distribution.

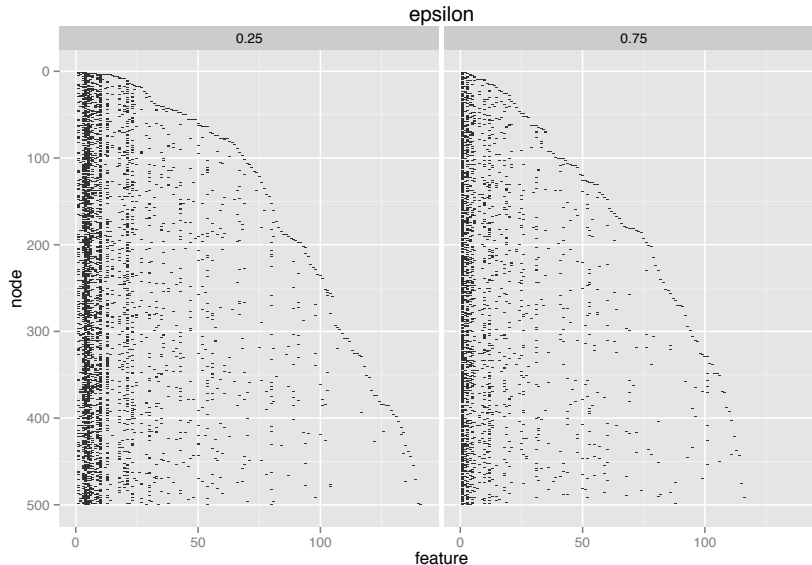


Figure 6: Here $n = 500$, $\alpha = 3$, $\beta = 0.5$ and the random variables R_i are uniformly distributed on the interval $[\varepsilon, 2 - \varepsilon]$ (so that $m_R = 1$ and $Var[R_i] = (1 - \varepsilon)^2/3$), for different values of ε ($\varepsilon = 0.25$ and $\varepsilon = 0.75$). The figure shows how ε affects the shape of Z .

3.3 Analysis of the ordering of the nodes

In a real application, we may content ourselves in finding not the realized fitness parameters themselves but rather their ordering, that is, the ordering of the nodes from larger to smaller values of the fitness parameter. To evaluate if we can at least extract values \hat{r}_i that respect this ordering, we decided to compare the drawn vector $\hat{\bar{r}}$ with the true realization \bar{r} by the use of Kendall's τ and some variants of it.

To keep track of the fact that, as said before, the first nodes contain more information than the last ones, we evaluated Kendall's τ not only on the whole vector but also on a short initial segment of size $k_n = n/2$ or $k_n = \sqrt{n}$. Besides this, we tried to use a variant of Kendall's τ (proposed in [35]), that we apply in two separate and different ways:

1. inducing a hyperbolic decay based on the position of the nodes – that is, weighting more the first (the oldest) nodes, and less the last (the youngest) ones;
2. inducing a hyperbolic decay based on the true realized values r_i – that is, assigning a higher weight to the nodes with a greater fitness parameter r_i .

The results of these measures are summarized in Table 1 for the experiment with the uniform distribution on an interval, in Table 2 for the experiment with

Considered nodes	Kendall's τ	τ weighted by position	τ weighted by value
$k_n = \lfloor \sqrt{n} \rfloor = 44$.281	.206	.463
$k_n = \frac{n}{2} = 1000$.229	.188	.337
$k_n = n = 2000$.150	.139	.155

Table 1: Comparing orderings induced by the true realization \bar{r} versus the extracted one \hat{r} in the case of the uniform distribution on the interval $[0.5, 1.5]$.

Considered nodes	Kendall's τ	τ weighted by position	τ weighted by value
$k_n = \lfloor \sqrt{n} \rfloor = 44$.676	.593	.713
$k_n = \frac{n}{2} = 1000$.586	.585	.625
$k_n = n = 2000$.438	.477	.434

Table 2: Comparing orderings induced by the true realization \bar{r} versus the extracted one \hat{r} in the case of the uniform discrete distribution on the two values $\{0.25, 1.75\}$.

Considered nodes	Kendall's τ	τ weighted by position	τ weighted by value
$k_n = \lfloor \sqrt{n} \rfloor = 44$.735	.762	.772
$k_n = \frac{n}{2} = 1000$.453	.516	.803
$k_n = n = 2000$.313	.402	.543

Table 3: Comparing orderings induced by the true realization \bar{r} versus the extracted one \hat{r} in the case of the normalized discrete Zipf's distribution with exponent 2 and 10 values.

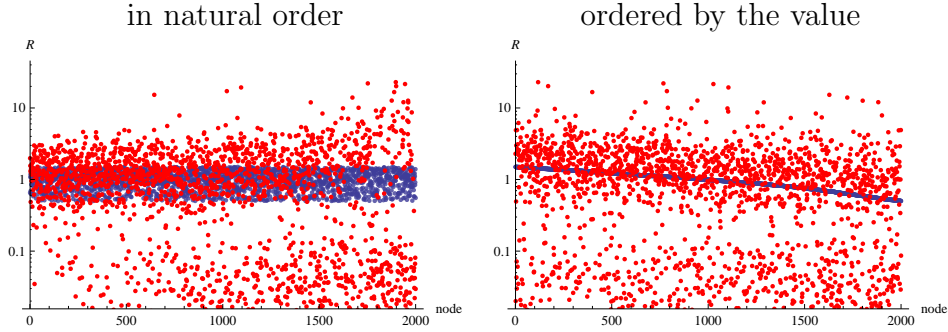


Figure 7: The extracted realization \widehat{r} (in red) versus the true realization \bar{r} (in blue), with two different orderings, in the case of uniform distribution on the interval $[0.5, 1.5]$. The empirical mean of the the first $\frac{n}{2}$ extracted values is 1.18.

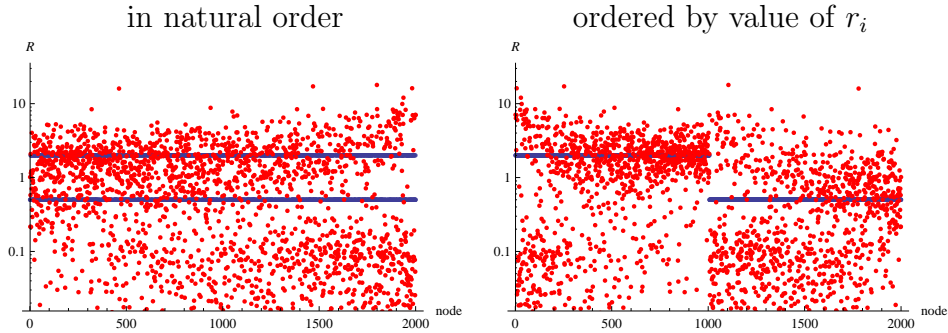


Figure 8: The extracted realization \widehat{r} (in red) versus the true realization \bar{r} (in blue), with two different orderings, in the case of the uniform discrete distribution on the two values $\{0.25, 1.75\}$. The empirical mean of the the first $\frac{n}{2}$ extracted values is 1.33.

the uniform discrete distribution on the two values $\{0.25, 1.75\}$, and in Table 3 for the discrete Zipf’s distribution with 10 values and exponent 2. The tables show that, although we are unable to reconstruct the actual realized values of the fitness parameters, our approach actually recovers some information about node ranking. As already seen before, the output of the Monte Carlo algorithm is better for the discrete cases.

4 From the attribute structure to the graph

We now extend the model to produce a graph out of the attribute structure (that may itself be latent and unknown). In general, we may assume that the presence of an edge between two nodes depends on the features that those nodes exhibit, but

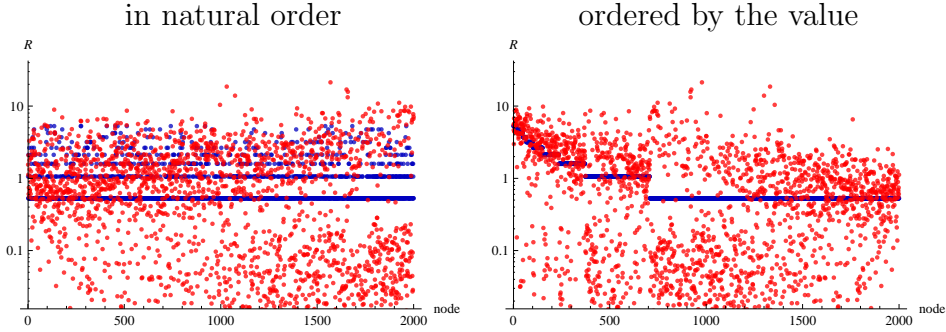


Figure 9: The extracted realization \widehat{r} (in red) versus the true realization \bar{r} (in blue), with two different orderings, for the normalized discrete Zipf's distribution with exponent 2 and 10 values. The empirical mean of the the first $\frac{n}{2}$ extracted values is 1.25.

there are many nuances to this idea and possible approaches.

In the sequel, we postulate that the connections are undirected (we omit self-loops, i.e., edges of type (i, i)) and we denote the adjacency matrix (symmetric by assumption) by A .

4.1 Feature/Feature probability model (FF)

In the first, basic model, we assume that the probability of having an edge (i, j) depends *solely* on the features that i and j possess; each pair of feature vectors that node i and node j exhibit contributes in tuning the edge probability. In other words, letting L_n be the total number of different features (i.e., columns of Z), we assume that there is a symmetric feature-feature influence matrix $\Xi = (\xi_{h,k})_{1 \leq h,k \leq L_n}$ that determines a node-node weight matrix W given by

$$W = Z \cdot \Xi \cdot Z^T$$

or, more explicitly,

$$w_{i,j} = \sum_{h,k} Z_{i,h} \xi_{h,k} Z_{j,k}.$$

The probability of the presence of an edge (i, j) , then, depends monotonically on $w_{i,j}$. The choice of Ξ determines different relations between features and edge probabilities. If $\xi_{h,k} > 0$ (resp., $\xi_{h,k} < 0$), then the simultaneous presence of attributes h and k increases (resp., decreases) the edge probability; if $\xi_{h,k} = 0$, the simultaneous presence of attributes h and k does not affect the edge-probability. In particular, if $\xi_{h,k} = 0$ for $h \neq k$, then the edge-probability is affected only by the presence of the same attributes in both nodes (positively or negatively affected depending on the sign of $\xi_{h,h}$).

The actual probabilities are computed as some function applied to the corresponding weight; i.e., some monotone function $\Phi : \mathbf{R} \rightarrow [0, 1]$ is fixed and, for $1 \leq j < i \leq n$,

$$P(A_{i,j} = 1|Z) = \Phi(w_{i,j}) = \Phi \left(\sum_{h,k} Z_{i,h} \xi_{h,k} Z_{j,k} \right). \quad (4.1)$$

4.2 Feature/Feature+BA probability model (FFBA)

A variant of the feature/feature (FF) probability model takes into account the fact that some edges exist independently of the features that the involved nodes exhibit, but they are there simply because of the popularity of a node, as in the traditional “preferential attachment” model by Barabási and Albert [3]. To take this into consideration, instead of using (4.1), we rather define for $1 \leq j < i \leq n$

$$P(A_{i,j} = 1|Z, D_j(i-1), m(i-1)) = \delta \Phi \left(\sum_{h,k} Z_{i,h} \xi_{h,k} Z_{j,k} \right) + (1-\delta) \frac{D_j(i-1)}{2m(i-1)}, \quad (4.2)$$

where $D_j(k)$ and $m(k)$ are, respectively, the degree of node j and the overall number of edges just after node k was added. The parameter δ controls the mixture between the pure feature/feature model and the preferential-attachment model (degenerating to the first one when $\delta = 1$, and to the second one when $\delta = 0$).

4.3 Feature/feature+JR probability model (FFJR)

Jackson and Rogers [20] observed that preferential attachment can be induced also injecting a “friend-of-friend” approach in the creation of edges. Their behavior can be mimicked in our model as follows: we first generate a graph with adjacency matrix A' using the pure FF model, i.e., letting

$$P(A'_{i,j} = 1|Z) = \Phi(w_{i,j}) = \Phi \left(\sum_{h,k} Z_{i,h} \xi_{h,k} Z_{j,k} \right).$$

After this, every node i looks at the set of the neighbors of its neighbors, according to A' . If this set is not empty, it then selects one node from the set uniformly at random; the resulting node is chosen as an “extra” friend of i with some probability $1 - \delta$ (for suitably chosen $\delta \in [0, 1]$). The adjacency matrix obtained in this way is A . Once more, if $\delta = 0$ we have $A = A'$ so we get back to the pure FF model.

5 Simulations for the graph structure

The purpose of this collection of experiments is to determine the topology of the graph generated with the models described above. We fix *a priori* the number of

nodes n and the (approximate) number of edges m (i.e., density) we aim at; then, every experiment consists essentially in two phases:

- generating an attribute matrix Z for n nodes (with certain values for the parameters α and β and with R_i uniformly distributed on the interval $[\varepsilon, 2 - \varepsilon]$ for a certain ε);
- building the graph according to one of the models described in Section 4.

The second phase needs to fix some further parameters: Ξ (the feature/feature influence matrix), the function Φ and, for the mixed models (FFBA and FFJR), the parameter δ .

For the sake of simplicity, throughout this section, we assume that $\Xi = I$ and we take Φ as a sigmoid function given by

$$\Phi(x) = \frac{1}{e^{K(\vartheta-x)} + 1}.$$

In other words, the existence of an edge (i, j) depends simply on the number of features that i and j share (this is an effect of choosing $\Xi = I$). More features induce larger probability: the sigmoid function smoothly increases (from 0 to 1) around a threshold ϑ , and $K > 0$ controls its smoothness; when $K \rightarrow \infty$ we obtain a step function and edges are chosen deterministically based on whether the two involved nodes share more than ϑ features or not.

In the experiments, we fix K and determine ϑ on the basis of the desired density of the graph (or, equivalently, the desired number of edges m); in practice⁴, this is obtained by solving numerically the equation

$$E \left[\sum_{1 \leq j < i \leq n} A_{i,j} \right] = \sum_{1 \leq j < i \leq n} \Phi \left(\sum_{h,k} Z_{i,h} \xi_{h,k} Z_{j,k} \right) = m$$

for the indeterminate ϑ (using, for example, Newton's method). Since $\Xi = I$ the equation in fact simplifies into

$$\sum_{1 \leq j < i \leq n} \frac{1}{e^{K(\vartheta - \sum_h Z_{i,h} Z_{j,h})} + 1} = m.$$

With these assumptions, every experiment depends on the parameters used for generating Z (i.e., α , β and ε), on K (that controls the smoothness of the sigmoid function) and on δ (for the mixed models). In the graphs produced by each simulation, we took into consideration the degree distribution, the percentage of reachable pairs (i.e., the fraction of pairs of nodes that are reachable) and the distribution of

⁴The described method needs some (obvious) adjustments when applied to the mixed models, to take into account the edges generated by preferential attachment.

distances (lengths of shortest paths); the latter data are computed using a probabilistic algorithm [32].

Some of the results obtained (for $n = 2000$ and⁵ $m = 4000$) for the FF model are shown in Figure 10. For those experiments, the underlying attribute matrix is generated with $\beta = 0.75$ and R_i uniformly distributed on the interval $[0.75, 1.25]$; we compare $\alpha = 3$ (resulting in ≈ 1200 features) with $\alpha = 10$ (≈ 4000 features). Results regarding mixed models are reported in Figure 11.

The properties of the obtained graphs can be summarized as follows:

- the pure FF model exhibits a behavior that strongly depends on the smoothness parameter K (see Fig. 10):
 - for $K = 1$, the degree distribution is power-law only when α is large (e.g., $\alpha = 10$), whereas the distribution is often non-monotonic for smaller α 's, especially on large graphs; the fraction of reachable pairs is quite large (between 40% and 90%);
 - for $K = 4$, degrees are always distributed as a power-law (with exponents around 3), but the graph becomes largely disconnected (the reachable pairs are never more than 20%): this is because nodes with the same degree tend to stick together (assortativity), forming a highly connected component and leaving the remaining nodes isolated;
 - for $K \rightarrow \infty$, the power-law distribution of degrees is even more clear-cut, but the number of reachable pairs becomes smaller (no more than 10%); the exponent of the power-law distribution depends on α , with larger α 's yielding larger absolute values of the (negative) exponents;
- the FFBA model (see Fig. 11) increases slightly the number of reachable pairs in all cases; the shape of the power-law distribution is essentially unchanged with respect to the pure FF model;
- finally, for the FFJR model (see Fig. 11) we observe a reduced connectivity; this is due to holding the expected number of edges as a constant, while devoting some of them to closing triangles – an operation that cannot increase connectivity. The degree distribution seems closer to a power-law with respect to the pure FF model.

6 A real dataset

We considered a dataset of scientific papers⁶ (originally released as part of the 2003 KDD Cup) consisting of 27 770 papers from the “High energy physics (the-

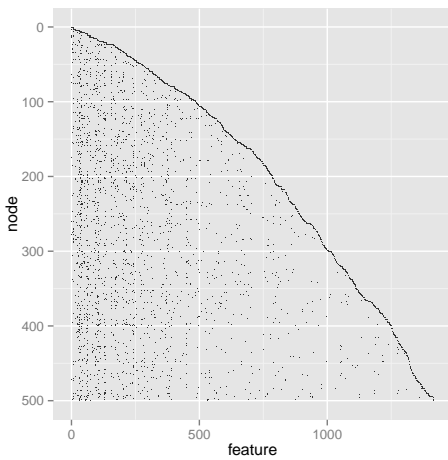
⁵We observed absolutely analogous phenomena also for larger and denser networks; we hereby report only the smaller case for the sake of readability of the pictures.

⁶The dataset is available within the SNAP (Stanford Large Network Dataset Collection) at <http://snap.stanford.edu/data/cit-HepTh.html>.

ory) arXiv” database. For each paper (node), we considered as features the words appearing in its title and abstract, excluding those that are dictionary words⁷. The papers are organized in order of publication date.

In Figure 12a the reader can see a fragment of the attribute matrix (for the first 500 nodes and the features they exhibit).

The overall number of features is 21 933, with a matrix density of $0.35 \cdot 10^{-3}$ (there are 214 510 ones in the matrix). The estimated values of α' and β are 15.038 and 0.671, respectively. In particular, we recall that β is the power-law exponent of the asymptotic behavior of L_n , i.e. the overall number of distinct attributes. We show the estimate for this real case in Figure 12b. A reconstruction of the ordering of the nodes according to their fitness parameter values is possible, but we lack any ground truth to compare it to.



(a) The first 500 rows (nodes) of the attribute matrix.



(b) Correspondence between the parameter β and the power-law exponent of L_n , as a function of n . The estimate of β is the slope of the regression line.

Figure 12: Analysis of the `cit-HepTh` dataset.

We conclude this section with a comparison between the graph produced by the FF model using as the underlying matrix the attribute matrix of the `cit-HepTh` dataset and the corresponding (symmetrized) citation graph. After some experiments, we observed that we can obtain a good fit with $K = 2.5$, that produces a quite similar degree and distance distribution (see Figure 13). It is striking to observe that the two graphs have such a strong similarity in their topology, albeit having positively no direct relation with each other (in one case the edges represent citations, in the other they were obtained by the model basing on the textual similarity of their abstracts!).

⁷According to the Unix `words` dictionary.

7 Conclusions

In this paper we introduce and study a network model that combines two features:

1. Behind the adjacency matrix of a network there is a *latent attribute structure* of the nodes, in the sense that each node is characterized by a number of features and the probability of the existence of an edge between two nodes depends on the features they share.
2. Not all nodes are equally successful in transmitting their own attributes to the new nodes (*competition*). Each node n is characterized by a random fitness parameter R_n describing its ability to transmit the node's attributes: the greater the value of the random variable R_n , the greater the probability that a feature of n will also be a feature of a new node, and so the greater the probability of the creation of an edge between n and the new node. Consequently, a node's connectivity does not depend on its age alone (so that also "young" nodes are able to compete and succeed in acquiring links).

Our work has different merits: firstly, we propose a simple model for the latent bipartite "node-attribute" network, where the role played by each single parameter is straightforward and easy to interpret: specifically, we have the two parameters, α and β , that control the number of new attributes each new node exhibits (in particular, $\beta > 0$ tunes the power-law behavior of the total number of distinct observed features); whereas the fitness parameters R_i impact on the probability of the new nodes to inherit the attributes of the previous nodes. Secondly, unlike other network models based on the standard Indian Buffet Process, we take into account the aspect of competition and, like in [5], we introduce random fitness parameters so that nodes have a different relevance in transmitting their features to the next nodes; finally, we provide some theoretical, as well experimental, results regarding the power-law behavior of the model and the estimation of the parameters. By experimental data, we also show how the proposed model for the attribute structure naturally leads to a complex network model.

The comparison with real datasets is promising: our model seems to produce quite realistic attribute matrices while at the same time capturing most local and global properties (e.g., degree distributions, connectivity and distance distributions) real networks exhibit.

Some possible future developments are the following. First, we could introduce another parameter $c \geq 0$ in the model of the node-attribute bipartite network so that the inclusion probabilities are

$$P_n(k) = \frac{\sum_{i=1}^n R_i Z_{i,k}}{c + \sum_{i=1}^n R_i}$$

(we now have $c = 0$): the bigger c , the smaller the inclusion probabilities and so the sparser the attributes. This can allow to obtain attribute matrices that are sparser.

To this purpose, we note that the proofs of the theoretical results regarding the estimation of α and β do not change and so, for this aspect, we have no problem. The problem is, instead, in the fact that we have an additional parameter to estimate.

Second, a possible variant of the feature/feature (FF) model is to consider, for each incoming new node i , a feature/feature influence matrix $\Xi(i)$ which depends on i : for instance, a diagonal matrix with

$$\xi_{k,k}(i) = \frac{1}{\sum_{\ell=1}^{i-1} Z_{\ell,k}}$$

so that the edge-probability is smaller as the number of nodes with k as a feature is larger.

Acknowledgments

Paolo Boldi and Corrado Monti acknowledge the EU-FET grant NADINE (GA 288956). They also would like to thank Andrea Marino for useful discussions.

Irene Crimaldi acknowledges support from CNR PNR Project “CRISIS Lab”. Moreover, she is a member of the Italian group “Gruppo Nazionale per l’Analisi Matematica, la Probabilità e le loro Applicazioni (GNAMPA)” of the Italian Institute “Istituto Nazionale di Alta Matematica (INdAM)”.

References

- [1] William Aiello and Fan Chung. Random evolution in massive graphs. In *Proc. of the 42Nd IEEE Symposium on Foundations of Computer Science, FOCS '01*, pages 510–, Washington, DC, USA, 2001. IEEE Computer Society.
- [2] Edoardo M. Airoldi, David M. Blei, Stephen E. Fienberg, and Eric P. Xing. Mixed membership stochastic blockmodels. *J. Mach. Learn. Res.*, 9:1981–2014, June 2008.
- [3] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286:509–512, October 1999.
- [4] Federico Bassetti, Irene Crimaldi, and Fabrizio Leisen. Conditionally identically distributed species sampling sequences. *Advances in Applied Probability*, 42(2):433–459, 06 2010.
- [5] Patrizia Berti, Irene Crimaldi, Luca Pratelli, and Pietro Rigo. Central limit theorems for an indian buffet model with random weights. *The Annals of Applied Probability*, 2014. (forthcoming). Currently available on http://imstat.org/aap/future_papers.html and on arXiv (1304.3626, 2013).
- [6] Ginestra Bianconi and Albert-László Barabási. Competition and multiscaling in evolving networks. *EPL (Europhysics Letters)*, 54(4):436, 2001.

- [7] Ronald L. Breiger. The Duality of Persons and Groups. *Social Forces*, 53(2):181–190, 1974.
- [8] Tamara Broderick, Michael I. Jordan, and Jim Pitman. Beta processes, stick-breaking and power laws. *Bayesian Analysis*, 7(2):439–476, 06 2012.
- [9] George Casella and Edward I George. Explaining the gibbs sampler. *The American Statistician*, 46(3):167–174, 1992.
- [10] Aaron Clauset, Cosma Rohilla Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Rev.*, 51(4):661–703, November 2009.
- [11] P. Erdős and A. Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.*, 5:17–61, 1960.
- [12] Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 6(6):721–741, 1984.
- [13] Charles J. Geyer and Minnesota Univ. (Minneapolis School Of Statistics). *Markov Chain Monte Carlo Maximum Likelihood*. Defense Technical Information Center, 1992.
- [14] Zoubin Ghahramani. Bayesian nonparametrics and the probabilistic approach to modelling. *Philosophical Transactions of the Royal Society A*, 371(20110553), 2012.
- [15] Thomas L. Griffiths and Zoubin Ghahramani. Infinite latent feature models and the indian buffet process. In *Advances in Neural Information Processing Systems*, pages 475–482. MIT Press, 2005.
- [16] Thomas L. Griffiths and Zoubin Ghahramani. The indian buffet process: An introduction and review. *J. Mach. Learn. Res.*, 12:1185–1224, 2011.
- [17] Peter Hall and C. C. Heyde. *Martingale Limit Theory and Its Application*. Academic Press, New York, NY, 1980.
- [18] Peter D. Hoff. Multiplicative latent factor models for description and prediction of social networks. *Computational and Mathematical Organization Theory*, 15(4):261–272, 2009.
- [19] Matthew O. Jackson. *Social and Economic Networks*. Princeton University Press, Princeton, NJ, USA, 2008.
- [20] Matthew O. Jackson and Brian W. Rogers. Meeting strangers and friends of friends: How random are social networks? *American Economic Review*, 97(3):890–915, 2007.

- [21] John F. Charles Kingman. Completely random measures. *Pacific Journal of Mathematics*, 21(1):59–78, 1967.
- [22] Ravi Kumar, Jasmine Novak, and Andrew Tomkins. Structure and evolution of online social networks. In *Proc. of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, pages 611–617, New York, NY, USA, 2006. ACM.
- [23] Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, D. Sivakumar, Andrew Tomkins, and Eli Upfal. Stochastic models for the web graph. In *Proc. of the 41st Annual Symposium on Foundations of Computer Science*, FOCS '00, pages 57–, Washington, DC, USA, 2000. IEEE Computer Society.
- [24] Matthieu Latapy, Clémence Magnien, and Nathalie Del Vecchio. Basic notions for the analysis of large two-mode networks. *Social Networks*, 30(1):31 – 48, 2008.
- [25] Silvio Lattanzi and D. Sivakumar. Affiliation networks. In *Proc. of the Forty-first Annual ACM Symposium on Theory of Computing*, STOC '09, pages 427–434, New York, NY, USA, 2009. ACM.
- [26] Kurt T. Miller, Thomas L. Griffiths, and Michael I. Jordan. Nonparametric latent feature models for link prediction. In Yoshua Bengio, Dale Schuurmans, John D. Lafferty, Christopher K. I. Williams, and Aron Culotta, editors, *In NIPS*, pages 1276–1284. Curran Associates, Inc., 2009.
- [27] Morten Mørup, Mikkel N. Schmidt, and Lars Kai Hansen. Infinite multiple membership relational modeling for complex networks. *CoRR*, abs/1101.5097, 2011.
- [28] Konstantina Palla, David A. Knowles, and Zoubin Ghahramani. An Infinite Latent Attribute Model for Network Data. In *Proc. of the 29th International Conference on Machine Learning*, pages 1–8, 2012.
- [29] Jim Pitman. *Combinatorial stochastic processes*, volume 1875 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2006.
- [30] Purnamrita Sarkar, Deepayan Chakrabarti, and Michael I. Jordan. Nonparametric link prediction in dynamic networks. In *Proc. of the 29th International Conference on Machine Learning*, pages 1–8, 2012.
- [31] Tom A.B. Snijders and Krzysztof Nowicki. Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification*, 14(1):75–100, 1997.

- [32] Sadagopan Srinivasan, Krithi Ramamritham, Arun Kumar, M.P. Ravindra, Elisa Bertino, and Ravi Kumar, editors. *HyperANF: approximating the neighbourhood function of very large graphs on a budget*. ACM, 2011.
- [33] Yee W. Teh and Dilan Gorur. Indian buffet processes with power-law behavior. In Y. Bengio, D. Schuurmans, J.D. Lafferty, C.K.I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1838–1846. Curran Associates, Inc., 2009.
- [34] Romain Thibaux and Michael I. Jordan. Hierarchical beta processes and the indian buffet process. In *Proc. 11th Conference on Artificial Intelligence and Statistics (AISTAT)*, pages 1–8, Puerto Rico, 2007.
- [35] Sebastiano Vigna. A weighted correlation index for rankings with ties. *CoRR*, abs/1404.3325, 2014.

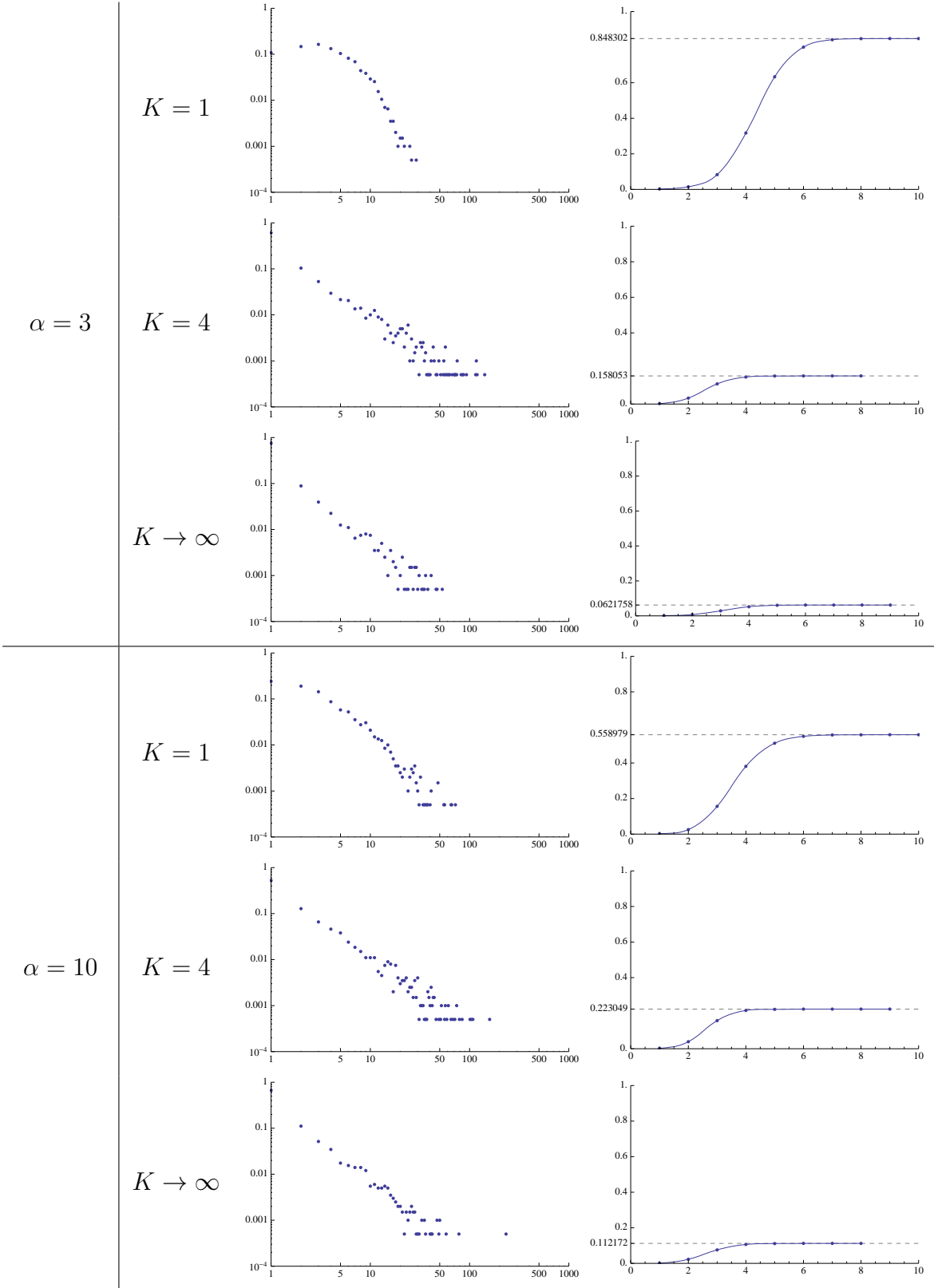


Figure 10: Properties of graphs generated by the FF model. We show the degree distribution in a log-log plot and the fraction of pairs at distance at most k ; in the latter, we highlight the peak value (fraction of mutually reachable pairs). 33

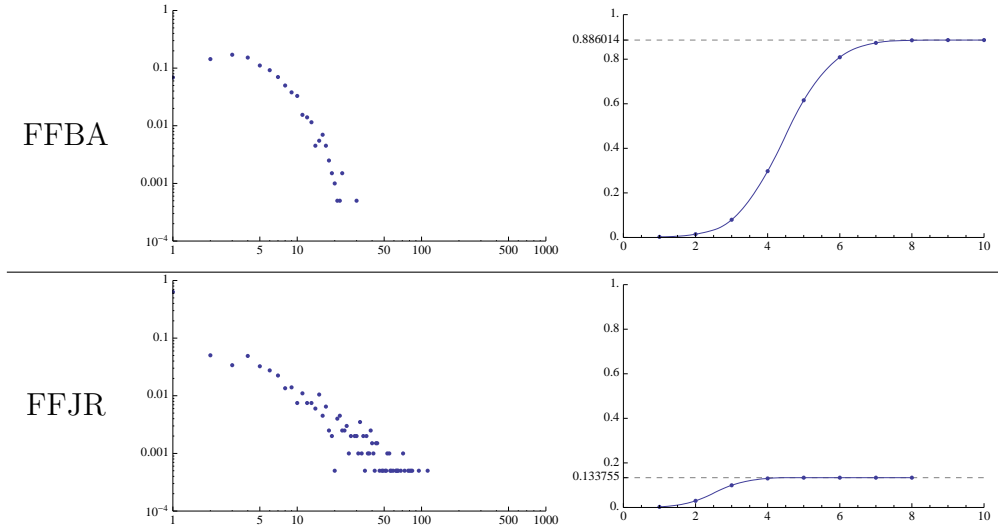


Figure 11: Properties of graphs generated by mixed models with $K = 1$ and $\delta = 0.75$. We show the degree distribution in a log-log plot and the fraction of pairs at distance at most k ; in the latter, we highlight the peak value, indicating how many pairs of nodes are mutually reachable. The parameters of the underlying attribute-matrix model are $\alpha = 3$ and $\beta = 0.75$ and the R_i 's are uniformly distributed on $[0.75, 1.25]$.

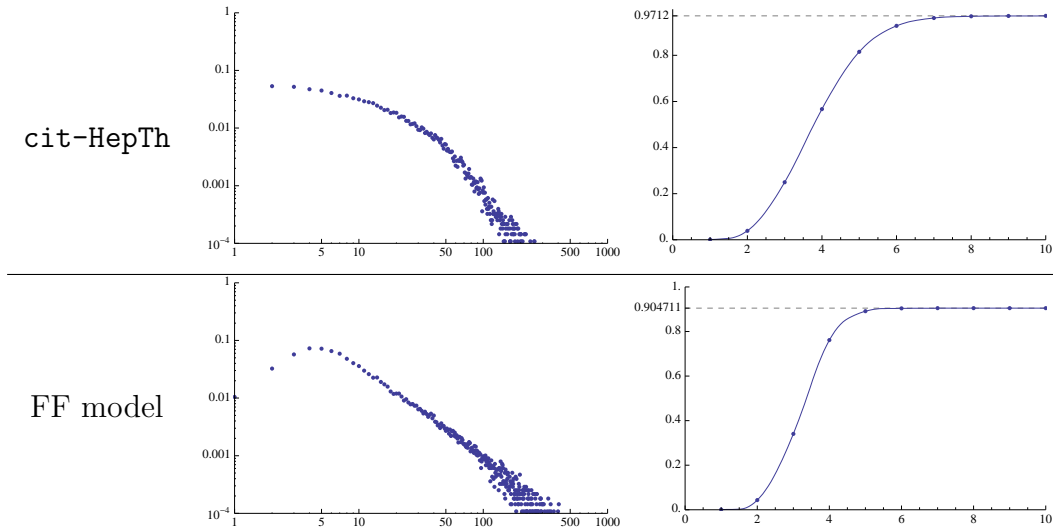


Figure 13: Comparison of the *cit-HepTh* dataset *versus* a graph generated by the FF model applied to the real feature matrix. We show the degree distribution in a log-log plot, and the fraction of pairs at distance at most k ; in the latter, we highlight the peak value, indicating how many pairs of nodes are mutually reachable.