A parametric approach to information filtering in complex networks: The Pólya filter

Riccardo Marcaccioli$^{1,2}$ and Giacomo Livan$^{1,2}$

$^1$Department of Computer Science, University College London, 66-72 Gower Street, London WC1E 6EA, UK
$^2$Systemic Risk Centre, London School of Economics and Political Sciences, Houghton Street, London WC2A 2AE, UK

The ever increasing availability of data demands for techniques to extract relevant information from complex interacting systems, which can often be represented as weighted networks. In recent years, a number of approaches have been proposed to extract network backbones by assessing the statistical significance of links against null hypotheses of random interaction. Yet, it is well known that the growth of most real-world networks is highly non-random, as past interactions between nodes typically increase the likelihood of further interaction. Here, we propose a network filtering methodology based on a family of null hypotheses that can be calibrated to assess which links are statistically significant with respect to a given network’s own heterogeneity. We design such family of null hypotheses by adapting the Pólya urn, a simple one-parameter combinatorial model driven by a self-reinforcement mechanism, to a network setting. We provide a full analytical description of our filter, and show that it retains or discards links based on a non-trivial interplay between their own local importance and the importance of the nodes they belong to. We prove that the widely used disparity filter can be recovered as a large-strength approximation for a specific value of the Pólya filter’s parameter, and illustrate our findings by applying the filter to two large network datasets.

I. INTRODUCTION

A vast number of complex interacting systems can be represented as networks [1]. Over the last 20 years, Network Science has been successfully applied in a wide range of disciplines, from Biology to Finance and the Social Sciences [2–6]. One of the main reasons behind such a success is that oftentimes network representations of seemingly very diverse systems share a number of common characteristics. A recurrent feature of several natural and social networks is the lack of a typical scale [2–7], i.e., the marked heterogeneity of major structural features such as the degree or strength distributions.

Understanding which nodes and links represent a set of structurally relevant interactions can be of crucial importance to obtain parsimonious descriptions of complex networks, and, indeed, has contributed to shed light on the functioning of a variety of systems, ranging from biological [8–9], social [10–11], financial [12] or even literature-related [13–14] systems. Furthermore, the size and, in some cases, the density of several real-world networks often prevent any meaningful visualization, and represent a major obstacle for clustering algorithms, which typically work well only with sparse systems [15,16]. Because of such challenges, a number of approaches to extract relevant information from complex networks have been developed over the years. Naturally, any filtering technique hinges on a definition of what type of information represents a signal as opposed to noise. As a result, the network backbones obtained through different filtering techniques carry different meanings and highlight different properties.

Early approaches to filtering focused on proximity networks, and relied on retaining interactions fulfilling some topological constraints. A seminal example of this kind of approach is the minimum spanning tree [17], which selects the tree with the highest total strength embedded in a network. Less constrained generalizations of such method are the planar maximally filtered graphs [18] and the triangulated maximally filtered graphs [19], which reduce topological complexity by forcing the embedding of network backbones on a surface.

Most of the methodologies initially proposed to filter information in weighted networks largely relied on discarding all links whose weights are below a certain global threshold [20–23], leading to backbones not reflecting the multiscale nature of the underlying network [24]. This issue has been addressed by a different class of techniques, which resort to hypothesis testing in order to assess the statistical significance of each link in a network. The disparity filter [25], which arguably represents one of most widely used filtering techniques, falls under this category, and relies on a null hypothesis of uniform distribution of a node’s strength over its links. Such a method has been adopted as one of the main benchmarks against which the efficiency of novel filtering techniques has been tested [26–29].

More recently, a procedure based on a null hypothesis of random connectivity (encoded as the urn problem described by the hypergeometric distribution) has been put forward [30–32]. Since such a procedure was the first to systematically make use of the Bonferroni multivariate correction [33] in the context of network filtering, whereas previous studies naively tested all links in a network against a univariate threshold, we shall refer to it as the Bonferroni filter.

Both the disparity and the Bonferroni filters provide a top-down approach based on well defined null hypotheses, against which all links in a network are tested in-
individual. While this certainly presents advantages in terms of convenience, at the same time it can lead to a lack of flexibility, as different networks may display different levels of heterogeneity, to which a “one-fits-all” null hypothesis cannot adapt. Furthermore, being based on null hypotheses of partially random interactions, both filters systematically tend to identify as statistically significant most links associated with large weights. Yet, the presence of “heavy” links in a heterogeneous networked system is the norm, rather than the exception, as past interactions naturally breed further interaction in a variety of natural and social systems [34, 35].

Here, we propose a filtering methodology based on a null hypothesis designed to respond to the specific heterogeneity of a network. We shall do so through a statistical test based on the Pólya urn, a well known combinatorial problem driven by a self-reinforcement mechanism according to which the observation of a certain event increases the probability of further observing it. Such a mechanism is governed by a single parameter $a$, which allows to tune the null hypothesis’ tolerance to heterogeneity, and to study a continuous family of network backbones $P_a$.

In the following, we shall detail how the Pólya filter works, both from an intuitive standpoint and by providing a full analytical characterization of the family of backbones it generates. In doing so, we shall show how the disparity filter can be recovered, with very good approximation, as a special case of the Pólya filter for $a = 1$.

II. RESULTS

A. The Pólya Filter

In the classic Pólya urn problem, we are given an urn containing $B_0$ black balls and $R_0$ red balls. We randomly draw a ball from the urn, we observe its colour and put it back in the urn together with a new balls of the same colour. When this process is repeated $n$ times, the probability of observing $x$ red balls follows the Beta-Binomial distribution [29] with probability mass function $P(x \mid n, \alpha, \beta) = \binom{n}{x} B(x + \alpha, n - x + \beta)/B(\alpha, \beta)$, where $B$ denotes the beta function and $\alpha = R_0/a, \beta = B_0/a$. In the following, we shall adapt this situation to a network setting.

Let us denote the $N \times N$ symmetric adjacency matrix of an undirected weighted network with $N$ nodes as $W$. An entry $w_{ij} \in \mathbb{N}$ of such a matrix is the weight associated with the link connecting nodes $i$ and $j$, and $w_{ij} = w_{ji} = 0$ when there is no connection between $i$ and $j$. The degree $k_i = \sum_{j=1}^{N} \mathbb{1}(w_{ij})$ (where $\mathbb{1}$ denotes the indicator function) quantifies the number of connections between a node $i$ and other nodes in the network, while $s_i = \sum_{j=1}^{N} w_{ij}$ denotes the strength of a node $i$, which is a measure of its activity in the network.

With the above notation, we can now rewrite the Pólya urn problem in network terms. Assume we are interested in assessing the statistical significance of a certain weight $w$ falling on one of the links of a node with degree $k$ and total strength $s$. Following the above example, we can think of this as a drawing process from a Pólya urn with 1 red ball and $k - 1$ black balls initially, where we want to measure the probability of drawing $w$ red balls in $s$ attempts. Such a probability reads

$$P(w \mid k, s, a) = \binom{s}{w} \frac{B\left(\frac{1}{a} + w, \frac{k-1}{a} + s - w\right)}{B\left(\frac{1}{a}, \frac{k-1}{a}\right)}.$$

The above equation fully describes our class of null hypotheses. We shall assume that a node distributes the weights on its links following a Pólya process whose reinforcement mechanism is governed by the parameter $a$. The rationale of such assumption lays in the flexibility introduced by such a parameter, which naturally captures situations where the more two nodes have interacted, the more further interactions between them become likely. In Fig. 1 we provide a sketch of the Pólya process adapted to a network setting.

Eq. (1) allows to compute the following $p$-values for a link of weight $w$:

$$\pi_p(w \mid k, s, a) = 1 - \sum_{w=0}^{w-1} P(x \mid k, s, a) = B\left(\frac{k-1}{a} + s - w, w + \frac{1}{a}\right) / (s + 1)B\left(\frac{1}{a}, \frac{k-1}{a}\right) \times 3F_2\left[1, w + 1, -s + w \atop \frac{1}{a}, -k - 1 \atop w + 1, -s + w\right].$$

where $B$ is the Beta function, and $3F_2$ denotes the generalized hypergeometric function. Once the value of the
The free parameter \( a \) has been set, two \( p \)-values can be assigned to the weight of each link in the network by applying Eq. (2) from the viewpoint of the two nodes it connects. The statistical significance of a weight is then assessed by comparing its associated \( p \)-values with a significance level. Since such a procedure involves testing all links in a network, this requires setting a univariate significance level \( \alpha \), and applying a multiple hypothesis test correction. The two main options available in this respect are the Bonferroni and the false discovery rate (FDR) corrections. The benefits and limitations of the two methods have been largely debated, and choosing between them essentially boils down to the type of statistical error one is more inclined to accept. The Bonferroni correction is much stricter than the FDR and typically ensures very high precision, leading to a low probability of accepting false positives, at the cost of a potentially low accuracy, i.e., of rejecting true positives. Following [30], in this work we shall adopt the Bonferroni correction criterion cannot be met with such a probability.

We shall now argue that the Pólya network backbone monotonically shrinks when the parameter \( a \) is increased, i.e.,

\[
 w \in \mathcal{P}_{a_2} \Rightarrow w \in \mathcal{P}_{a_1}, \text{ for } a_1 \leq a_2 .
\]

In other words, the largest Pólya set is the one corresponding to \( a = 0 \), and increasing \( a \) progressively removes links from this set.

In order to understand how the link removal process works, let us define the ratio:

\[
r = \frac{w}{s}.
\]

The rationale for introducing the above ratio will be justified later. For the moment, let us remark that it can be intuitively interpreted as a measure of weight concentration \((w/s)\) for links belonging to a node with degree \( k \).

From Eq. (1), the expected mean and variance under the assumed null hypothesis can be computed:

\[
\mu_r = \mathbb{E}[r] = 1,
\]

\[
\sigma_r^2(k, s, a) = \mathbb{E}[(r - \mu_r)^2] = \frac{k - 1}{s} \frac{k + as}{a + k} .
\]

Let us suppose that, once a significance level has been set, the links included in the Pólya backbone are those whose value of the ratio \( r \) in Eq. (4) is such that \( \mu_r - b \sigma_r(k, s, a) < r < \mu_r + b \sigma_r(k, s, a) \), for a certain \( b > 0 \). As it can be seen from Eq. (5), the standard deviation \( \sigma_r(k, s, a) \) increases monotonically as a function of \( a \). This, in turn, means that for fixed degree \( k \) and strength \( s \) the range of the ratio \( r \) compatible with the null hypothesis becomes larger as \( a \) increases, therefore confirming the property in Eq. (3). In the top panels of Fig. 2 we illustrate these results on two network datasets (US airport network [40, 41] and World Input-Output Database [42], see Appendix A for a brief description).

The bottom panels in Fig. 2 further confirm that, indeed, higher values of \( r \) tend to be associated with a higher statistical significance (and that such significance, in turn, decreases with higher values of \( a \)), although this is not a strict relationship due to the dependency on \( k \) and \( s \) of the standard deviation in Eq. (5). The same considerations can be drawn by looking at the network visualizations in Fig. 3. Further light on such dependencies will be shed in the following section.

The content of Eqs. (3)-(5) may also be intuitively seen as follows. The null hypothesis generated by taking \( a = a_1 \) can approximately be considered the same as the one associated with \( a = a_2 > a_1 \) provided a larger confidence interval for the evaluation of the second null hypothesis is set.

C. Generalizing the disparity filter

Eq. (2) can be considerably simplified under the assumption \( s \gg k, w \gg 1 \). In this regime, the \( p \)-values the
that is known to be a limit case of the Beta-Binomial distribution as the number of draws goes to infinity.

The same Figure also shows the relationship between the two sets of $p$-values for other values of $a$, which sheds light on the different nature of the backbones obtained from the application of the two filters. Setting $a > 1$ increases the Pólya filter’s tolerance to heterogeneity with respect to the disparity filter’s, which, as outlined in the previous section, results in a larger number of links being compatible with the filter’s null hypothesis, i.e., in a more restrictive test. Ultimately, this produces a much smaller backbone than the disparity filter. Conversely, setting $a < 1$ sets a low tolerance for heterogeneity, which ensures that several links rejected by the disparity filter are instead included in the Pólya filter’s backbone.

D. The role of the $r$ ratio

If we further approximate Eq. (6) by expanding it around $w/s \approx 0$ we obtain

\[ \pi_p \approx \frac{e^{-\frac{5}{2}} \left( \frac{s}{a} \right) \frac{1}{2} - 1}{\Gamma \left( \frac{1}{a} \right)} \cdot \frac{1}{\Gamma \left( \frac{1}{2} \right)} , \]

where $r = kw/s$ was introduced in Eq. (1). This result justifies the soft dependence of the Pólya filter on the ratio $r$ and clarifies the results presented in the previous section.

The ratio $r$ couples a network’s local topology (through the degree $k$) to the activity of nodes (through the strength $s$ and weight $w$) in a non-trivial way. The soft dependence of the Pólya filter on such quantity is what ensures that the Pólya backbones (for reasonable values of $a$) retain the multiscale nature of the networks they are extracted from. Indeed, the assumptions under which the approximation in Eq. (8) is obtained (i.e., $s \gg k, w \gg 1$) inform us that node strength is the primary driver of statistical significance under the Pólya (and consequently the disparity) filter. Yet, regardless of the local relative importance of a particular link (measured in terms of relative strength $w/s$), a large enough degree will ensure a large value of the ratio $r$, which will ultimately lead to low $p$-values. Still, links can be retained in a Pólya backbone even if not attached to high-degree or high-strength nodes. This is shown in the bottom panels of Fig. 2, where one can see that there are substantial exceptions to the trend of Eq. (8), which are precisely induced by the interplay between topology and node activity captured by the ratio $r$.

E. Fixing the free parameter

As we argued, one of the benefits of the Pólya filter is its flexibility, in that it allows to explore the network backbones obtained when setting different levels of tolerance to heterogeneity, as quantified by the parameter $a$. 

FIG. 2: Role of the parameter $r$ (see Eq. (1)) in the Pólya backbone extraction process. **Top left:** Evolution of the minimum, maximum, and average value of $r$ computed in Pólya backbones for increasing values of $a$ with a univariate significance level $\alpha = 0.05$ in the US airport network. **Top right:** Same quantities computed in the WIOT network. **Bottom left:** Scatter plots of the $p$-values associated with each link in the US airport network against the corresponding value of the ratio $r$ for two different values of $a$ at a univariate significance level $\alpha = 0.05$. High values of $r$ are associated with $p$-values below the Bonferroni threshold $\alpha_B$ (solid black line), while the opposite is not always true. The black dashed lines illustrate the soft dependence on $r$ described by Eq. (5). **Bottom right:** Same plot for the WIOT network.
We devote this section to recommending possible criteria that would identify an optimal value $a^*$ of such a parameter. Clearly, the notion of optimality will strongly depend on the specific application being considered. Therefore, we will recommend three different criteria.

**Sweeping.** The Pólya filter’s monotonicity can be exploited to fix a desired level of sparsity of the resulting backbone with respect to the original network, and to identify the value $a^*$ that achieves it. Namely, as a consequence of the property in Eq. (3), the fraction of nodes, of edges, and of total strength retained in the Pólya backbones are all monotonically non-increasing functions of the parameter $a$. Hence, starting from $a = 0$, one can scan the backbone family $P_a$ for increasing values of $a$ until a desired level of sparsity has been reached (e.g., 5% of the nodes in the original network).

**Maximum likelihood.** Eq. (1) can be used to define a log-likelihood function, which in turn be shown to have a maximum (see Appendix D). By definition, such a value corresponds to the Pólya process whose self-reinforcement mechanism is the most likely to generate the network under study. Effectively, this amounts to identifying the value $a_{ML}$ corresponding to the “nullest” model in the Pólya family or, in other words, the Pólya process that best captures the heterogeneity of the network under consideration. We further convey this point in Appendix D by showing on synthetic networks that the maximum likelihood estimates of the parameter $a$ are indeed sensitive to changes in the network’s heterogeneity. As such, this criterion is particularly suited to applications where validating the backbone as a whole is a priority.

**Salience.** Lastly, we are going to propose an ad-hoc criterion based on a compromise between the information retained in a backbone and the information lost by filtering the network it is extracted from. We shall quantify the former in terms of salience $S(a)$, a recently proposed yet well established measure of link importance, which can be loosely defined as the fraction of weighted shortest-path trees a link participates in. Such a measure has been shown to account for both the topological position of a link and for the magnitude of its associated weight (somewhat in analogy to the quantity in Eq. (4)), and captures several essential transport properties. In Appendix D we show that, as $a$ increases, the links removed from Pólya backbones are generally those with a lower salience. As a result, the average salience $\langle S(a) \rangle$ retained in the backbones $P_a$ increases with $a$. 

![FIG. 3: Visualizations of network backbones obtained with the Pólya filter. Both visualizations have been created using a software for large networks analysis (FNA Platform v. 18.4.5). **Left:** Radial tree visualization of the largest connected component of the US Airports Network filtered using two different values of the parameter $a$ ($a = 1$ and $a = 2$). Links included in both backbones are drawn in purple, while links included only in the $a = 2$ backbone are drawn in grey. Link opacities are proportional to the corresponding values of $r$. **Right:** Bipartite visualization of a subset of the WIOT network backbones obtained for two different values of the parameter $a$ ($a = 1$ and $a = 2$). Links included in both backbones are drawn in purple, while links included only in the $a = 2$ backbone are drawn in grey. Link opacities are proportional to the corresponding values of $r$, while node sizes are proportional to their total strength. Both panels visually confirm the soft dependency of the filter on $a$ values of $a$.](image1)

![FIG. 4: **Left:** Comparison of the $\pi_D$ values computed by the disparity ($\pi_D$) and Pólya ($\pi_P$) filters computed for different values of $a$ in the US airport networks (at a univariate significance level $\alpha = 0.05$). Each region of the plot is colored depending on the significance of the two filters. Points in the blue (green) region correspond to links rejected (accepted) by both filters, while points in the purple (red) region correspond to links accepted only by the disparity (Pólya) filter. **Right:** Same plot for the WIOT network.](image2)
Measuring the quality of a backbone just in terms of average salience could lead, in most cases, to an excessive depletion of the network under study. This tendency can be contrasted by penalizing large differences between backbones and their original networks. We do so by introducing the two following optimality measures

\[ O_1 = J(W, P_a) \cdot \langle S(a) \rangle, \quad O_2 = F_n(a) \cdot \langle S(a) \rangle, \quad (9) \]

where we are weighting the average salience against the Jaccard similarity \( J(W, P_a) \) between the weights in the original network and those in the backbone, and against the fraction \( F_n(a) \) of nodes retained in \( P_a \), respectively. Fig. 5 shows the behavior of the above metrics as functions of \( a \) in the two networks we study. As it can be seen, both metrics achieve a maximum \( a^* \), which represents the optimal compromise between high salience and similarity with respect to the original network.

F. Heterogeneity and weight rescaling

In order to get a further grasp of how the Pólya filter responds to different levels of heterogeneity, in this section we analyse how a rescaling of all weights of a network by a factor \( c \) affects the Pólya backbones. Referring once again to (5), one can see that \( \partial \sigma^2(k,c,s,a)/\partial c < 0 \) for \( c > 0 \), i.e., the variance of the ratio \( r \) decreases monotonically upon rescaling. This, in turn, will typically lead to a larger number of links being validated by the Pólya filter and to larger backbones.

In order to understand which links are particularly affected by a rescaling, let us compute the limit:

\[ \lim_{c \to \infty} \frac{\sigma^2_r(k,c,s,a)}{\sigma^2_r(k,s,a)} = \frac{as}{k + as}. \quad (10) \]

Since the above limit approaches 1 when \( s \gg k \), it corroborates the previously made observation that links attached to high-strength nodes tend to be validated by the Pólya filter regardless of the degree. On the other hand, for a fixed level of strength \( s \) an increase in degree \( k \) causes the above ratio to decrease, which in turn means that the Pólya filter will increasingly tend to validate links attached to hubs in the network when a global rescaling is performed.

III. CONCLUSIONS

In the era of Big Data, information filtering methods are needed more than ever to handle the dazzling complexity of both social and natural networked systems. In this paper, we have proposed a novel technique based on the Pólya urn model to extract backbones of statistically relevant interactions between pairs of nodes in a network. In the network context, the parameter \( a \) tuning the Pólya model’s self-reinforcement mechanism effectively becomes a tolerance to a network’s heterogeneity. This, in turn, introduces an element of flexibility, which, to the best of our knowledge, other network filtering techniques do not provide.

Indeed, we have shown that the Pólya filter generates a continuous family of network backbones. Depending on the specific application, the null hypothesis underpinning the filter can be chosen so as to have a different tolerance to heterogeneity. The low-tolerance regime \((a < 1)\) corresponds to a rather loose filtering, suited to situations where the main goal is to filter out interactions that can be unquestionably identified as noise. On the other hand, the high-tolerance regime \((a > 1)\) corresponds to increasingly restrictive tests, where only links of substantial structural importance survive.

As we have shown, the link selection criterion underpinning the Pólya filter is based on the interplay between topology and the local relative importance of a link, quantified by the parameter \( r \). This, in turn, guarantees that the filter does not perform a naive link selection merely based on retaining high strength links connecting hubs, but instead ensures a non-trivial scanning of all the relevant scales of a network.
Appendix A: Data

In the following we provide a short description the datasets we employed to illustrate the Pólya filter.

World Input Output Database The Database contains yearly aggregate economic transactions, measured in millions of dollars, between the industrial sectors of different countries from 2000 to 2014. The database features transactions between 64 sectors in 45 countries \cite{42,46}. The resulting series of networks and their properties have been analyzed extensively in a number of studies \cite{47,49}. The dataset we are going to use in this paper is the 2014 network, which features 2,464 nodes and 738,374 edges.

US Airports network The dataset contains information on the flights between a number of US airports during the year 2017 \cite{50}. Each link represents a connection between airports, with the weight representing the number of passengers on all flights on that route in the given direction. The system contains 1151 airports and 20,580 different connections. The same network with data coming from year 2010 is analysed in \cite{40,41}.

Appendix B: The Pólya filter for directed weighted networks

Systems where the directionality of interactions cannot be neglected are usually described in terms of directed weighted networks \cite{2,4}. The difference between weighted directed and weighted undirected networks is that the activity of each node can be specified in terms of a single degree \(k\) associated with two \(p\)-values to each weight by assessing its statistical significance both as an incoming and as an outgoing link. For example, when testing as an outgoing link, Eq. (2) is easily generalized as (we drop all node indices to keep notation light)

\[
\pi_P(w \mid k_{out}, s_{out}, a) = \frac{B \left( k_{out} - \frac{1}{a} + s_{out} - w, w + \frac{1}{a} \right)}{(s_{out} + 1)B \left( \frac{k_{out} - 1}{a}, w + 1 \right)} \cdot \text{B}(s_{out} - w + 1, w + 1) \end{aligned} 
\]

with the replacements \(k_{out} \rightarrow k_{in}, s_{out} \rightarrow s_{in}\) for the test as an incoming link. Both \(p\)-values can be tested against the same univariate threshold \(\alpha\). A link is retained by the Pólya filter only when at least one of the two \(p\)-values is lower than \(\alpha\).

A link is kept only if at least one of the two \(p\)-values is lower than \(\alpha_B\). In the case where \(k_{out} = 1\), we keep the directed link connecting \(i\) and \(j\) only if \(\pi_P(w_{ij} \mid k_{in}^i, s_{in}^i, a) < \alpha_B\), and vice versa in the case \(k_{in}^j = 1\).

Appendix C: Approximating the Pólya \(p\)-value

In this section we explicitly show how the disparity filter can be recovered as a special case of the Pólya filter for \(\alpha = 1\). We start by rewriting the \(p\)-value associated with a weight \(w\) attached to a node with degree \(k\) and strength \(s\). For the sake of simplicity, we go back to the undirected case (Eq. (2)):

\[
\pi_P(w \mid k, s, a) = \frac{B \left( k - \frac{1}{a} + s - w, w + \frac{1}{a} \right)}{(s + 1)B \left( \frac{k - 1}{a}, k - \frac{1}{a} \right) B(s - w + 1, w + 1)} \cdot \text{B}(s - w + 1, w + 1) 
\]

In the following we will repeatedly simplify the above expression by making use of the zero-order Stirling approximation for the ratio of two Gamma functions:

\[
\frac{\Gamma[x + \alpha]}{\Gamma[x + \beta]} = x^{\alpha - \beta} \left( 1 + O \left[ \frac{1}{x} \right] \right) \approx x^{\alpha - \beta},
\]
which holds for $x \to \infty$.

We first take care of the hypergeometric function in Eq. (C1). We start by expanding it in terms of ratios of Gamma functions:

$$3F_2 \left[ \begin{array}{c} 1, w + \frac{1}{a} - s_i + w \\ w + 1, -\frac{k - 1}{a} - s + w \end{array} : 1 \right] = \sum_{n=0}^{\infty} \frac{\Gamma[-s + w + n]}{\Gamma[-s + w]} \frac{\Gamma\left[\frac{-k-1}{a} - s + w + 1\right]}{\Gamma\left[\frac{-k-1}{a} - s + w + 1 + n\right]} \frac{\Gamma\left[w + \frac{1}{a} + n\right]}{\Gamma\left[w + \frac{1}{a}\right]} \frac{\Gamma[w + 1]}{\Gamma[w + 1 + n]}.$$  \hspace{1cm} (C3)

We can simplify the last two terms in the above expression:

$$\frac{\Gamma\left[w + \frac{1}{a} + n\right]}{\Gamma[w + 1 + n]} \frac{\Gamma[w + 1]}{\Gamma[w + \frac{1}{a}]} \approx w^{\frac{1}{2} + n} \approx 1,$$

where we have assumed $w \gg \frac{1}{a}$. Putting this result back into Eq. (C3) gives:

$$3F_2 \left[ \begin{array}{c} 1, w + 1 + \frac{1}{a} - s + w + 1 \\ w + 2, -\frac{k - 1}{a} - s + w + 2 \end{array} : 1 \right] \approx 2F_1 \left[ \begin{array}{c} -s + w, 1 \\ -\frac{k - 1}{a} - s + w + 1 : 1 \right].$$  \hspace{1cm} (C4)

Eq. (C4) can be now further simplified by making use of the the Chu-Vandermonde identity $2F_1(-n; b; c) = \frac{(-b)_n}{(c)_n}$ (where $(\cdot)_n$ denotes the Pochhammer symbol), which gives:

$$2F_1 \left[ \begin{array}{c} -s + w, 1 \\ -\frac{k - 1}{a} - s + w + 1 : 1 \right] = \frac{s - w + \frac{k-1}{a}}{(k-1)/a}.$$  \hspace{1cm} (C5)

Putting Eq. (C5) back into Eq. (C1) and writing the Beta functions in Eq. (C1) as ratios of Gamma functions, allows to write Eq. (C1) as the product of the three following ingredients:

$$B \left[ \frac{k - 1}{a} + s - w, w + \frac{1}{a} \right] \left( s - w + \frac{k-1}{a} \right) = \frac{\Gamma\left[\frac{k-1}{a} + s - w + 1\right]}{\Gamma\left[\frac{s + 1}{a}\right]} \frac{\Gamma\left[w + \frac{1}{a}\right]}{\Gamma\left[w + 1\right]} \frac{\Gamma\left[\frac{w}{a}\right]}{\Gamma\left[\frac{w}{a} + \frac{1}{a}\right]}.$$  \hspace{1cm} (C6)

By matching Gamma functions in the numerators and denominators of the above ratios, and making use of the Stirling approximation (Eq. (C2)), we can then write down the $p$-value in Eq. (C1) as the product of the following quantities:

$$\frac{\Gamma\left[s - w + \frac{k-1}{a} + 1\right]}{\Gamma\left[s - w\right]} \approx (s - w)^{\frac{k-1}{a}} = s^{\frac{k-1}{a}} \left(1 - \frac{w}{s}\right)^{\frac{k-1}{a}}, \quad s - w \gg \frac{k-1}{a} + 1$$

$$\frac{\Gamma\left[w + \frac{1}{a}\right]}{\Gamma\left[w + 1\right]} \approx w^{\frac{1}{a} - 1}, \quad w \gg \frac{1}{a}, w \gg 1$$

$$\frac{\Gamma\left[s + 1\right]}{\Gamma\left[s + \frac{k}{a}\right]} \approx s^{\frac{k}{a} - \frac{1}{a}}, \quad s \gg \frac{k}{a}, s \gg 1$$

$$\frac{\Gamma\left[\frac{w}{a}\right]}{\Gamma\left[\frac{w}{a} + \frac{1}{a}\right]} \approx s^{\frac{k}{a} - \frac{1}{a}}, \quad k \gg a - 1,$$

where on each line we have written the approximations we made use of. Finally, we can put together the above expressions, which gives the result reported in Eq. 6.

$$\pi_P(w \mid k, s, a) \approx \frac{1}{\Gamma\left[\frac{k}{a}\right]} \left(1 - \frac{w}{s}\right)^{\frac{k-1}{a}} \left(\frac{w k}{s a}\right)^{\frac{1}{a} - 1},$$  \hspace{1cm} (C8)
FIG. 6: Scatter plots of the $p$-values obtained from the Pólya filters compared with the approximate expression in Eq. (C8) for different values of the parameter $a$.

from which the disparity filter’s $p$-value can be recovered for $a = 1$ (see Eq. (7)).

All the approximations that we are assuming are written in Eq. (C7). In Fig. 6 we show a comparison between the $p$-values obtained from the Pólya filter (Eq. (C1)) and the above expression in the two networks we consider. As it can be seen, the overall agreement is rather good, and larger values of $a$ improve the quality of the approximation, as it can be seen from Eq. (C7).

Appendix D: Maximum Likelihood Estimates

As a parametric approach, the Pólya filter lends itself to optimization procedures aimed at identifying the value of the parameter $a$ most suited to the particular network under study. As mentioned in the text, maximum-likelihood estimation (MLE) is a natural option to single out the “nulllest” model in the Pólya family for the network under consideration.

This can be achieved by solving

$$a_{ML} = \arg \max_{a \in [0, \infty)} \mathcal{L}(a; w) ,$$

(D1)

where $w$ denotes the sequence of weights in the network, and

$$\mathcal{L}(a; w) = \sum_{i,j=1}^{N} \log P(w_{ij} \mid s_i, k_i) = \sum_{i,j=1}^{N} \log \left[ \frac{s_i}{w_{ij}} B\left( \frac{1}{a} + \frac{k_i - 1}{a} + s_i - w_{ij} \right) \right]$$

(D2)

is the log-likelihood function associated with the probability of observing the particular weight sequence under a Pólya process with parameter $a$.

Solving the optimization problem in (D1) with the above function boils down to solving numerically the following equation:

$$\sum_{i,j=1}^{N} - (k_i - 1) \psi \left( \frac{k_i + a s_i - a w_{ij} - 1}{a} \right) + k_i \psi \left( \frac{k_i}{a} + s_i \right) + (k_i - 1) \psi \left( \frac{k_i - 1}{a} \right) - k_i \psi \left( \frac{k_i}{a} \right) - \psi \left( \frac{w_{ij} + 1}{a} \right) + \psi \left( \frac{1}{a} \right) = 0 ,$$

(D3)

where $\psi(x) = \Gamma'(x)/\Gamma(x)$ is the Polygamma function of order 0.

In Fig. 7 we report ML estimates obtained on synthetic networks. The networks employed in the top plot are characterised by a scale-free topology generated using the BA model [2] and a power-law weight distribution with tail exponent $\tau$. The optimal values $a_{ML}$ clearly show that ML estimates respond to the network’s heterogeneity, spanning almost three orders of magnitude ranging from values $a_{ML} \simeq 10$ in the presence of very strong heterogeneity ($\tau = 1.5$) to $a_{ML} \simeq 10^{-3} - 10^{-2}$ in the presence of mild heterogeneity. In the lower plot of Fig. 7 we report the ML
FIG. 7: ML estimates of the Pólya filter’s parameter $a$. In both cases, the networks are made of 3,000 nodes and have an average degree of 8. The error bars are 95% confidence intervals obtained through 200 different randomizations of both weights and topology. **Left:** ML estimates $a_{ML}$ for Barabasi-Albert networks with a power-law weight distribution with tail exponent $\tau$. **Right:** ML estimates $a_{ML}$ for Erdős-Rényi networks with uniform distribution of weights $U[1, \tau]$. 

estimates on Erdős-Rényi random graphs with a uniform weight distribution $U[1, \tau]$, with weights rounded to the nearest integer. As it can been, the estimates are much less sensitive to changes with respect to the previous case, with $a_{ML} \simeq 0.46$–$0.56$, which implies the *de facto* impossibility to discriminate even between substantially different models when no marked heterogeneity is present in their weight distributions.

**Appendix E: Relationship with salience**

Link salience is a recently introduced measure of link importance [15], based on the distance between nodes. Given the adjacency matrix $W$ of weighted directed network, where an element $w_{ij}$ represent the strength of the interaction between nodes $i$ and $j$, the salience is computed through the auxiliary distance matrix $D$ such that $d_{ij} = 1/w_{ij}$ if $w_{ij} > 0$ and 0 otherwise. Once $D$ is known, the salience of a connection $(i, j)$ can be obtained. For a fixed reference node $r$, the set of weighted shortest paths to all other nodes is called the shortest-path tree matrix $T(r)$, which collects the most effective routes from $r$ to the rest of the network. $T(r)$ is a symmetric $N \times N$ matrix $T(r)$ such that $t_{ij}(r) = 1$ if the link $(i, j)$ is part of at least one of the shortest paths starting from $r$ and $t_{ij}(r) = 0$ otherwise. Once all the possible $T(r)$ $r = 1, 2 \ldots N$ matrices have been calculated, the salience of a link $(i, j)$ can be computed as:

$$S_{ij} = \frac{1}{N} \sum_{r=1}^{N} t_{ij}(r).$$  

(E1)

For a large collection of complex networks, it has been found that the distribution of link salience exhibits a peculiar bimodal shape in the unit interval, with most links ending up with $S \approx 0$ or $S \approx 1$. As a result, salience could be used to extract a network backbone, as this would practically not be affected by any particular salience threshold.

Interestingly, the Pólya filter displays an empirical relationship with the salience. In both the WIOT and the US Airport network, we verify that, as we increase the parameter $a$, the filter has a tendency to retain links with higher salience. We show this in Fig. 8 by plotting the mean and the skewness of the link salience distribution in both networks computed only in the links retained in the Pólya backbones. As it can be seen, the mean increases (not necessary monotonically) while the skewness decreases as $a$ is raised.

The intuition behind this can be found once again in the ratio $r = kw/s$ (see Eq. (4)). Indeed, we have shown that links associated with a higher $r$ are typically assigned lower p-values by the Pólya filter. The same can be said for the salience, whose scores appear to have a positive and statistically significant rank correlation with the corresponding values of $r$: corr$(r, s) \approx 0.3$ in the US Airport network, and corr$(r, s) \approx 0.2$ in the WIOT network.
FIG. 8: Evidence that the Pólya filter selects links with higher salience. **Left:** Average salience progressively calculated only in the links included in the backbones: $\langle S \rangle_a = \frac{1}{l_a} \sum_{(i,j) \in P_a} S_{ij}$ where $l_a$ is the number of links in the backbone $P_a$. **Right:** Skewness of the salience distribution calculated only in the links included in the backbones: skew($S$) = skew_{(i,j) \in P_a} S_{ij} (S)$ where $l_a$ is the number of links in the backbone $P_a$.

[1] V. Latora, G. Russo, and V. Nicosia, *Complex networks principles, methods and applications* (Cambridge University Press, 2017).
[2] R. Albert and A.-L. Barabási, Reviews of Modern Physics 74, 47 (2002), ISSN 0034-6861, 0106096v1.
[3] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and H. C. van de Hulst, Phys. Rep. 424, 175 (2006), ISSN 0370-1573.
[4] N. MEJ., SIAM Rev 45, 167 (2003).
[5] A. Vespignani, M. Bathelemy, and A. Barrat, *Dynamical Processes on Complex Networks* (2008), ISBN 9780521879507, arXiv:1011.1669v3.
[6] R. N. Mantegna, E. H. Stanley, and N. A. Chriss, Physics Today 53, 70 (2000), ISSN 0031-9228.
[7] G. Caldarelli, OUP Catalogue p. 328 (2007), ISSN 00368733.
[8] X. Zhou, J. Menche, A.-L. Barabási, and A. Sharma, Nature communications 5, 4212 (2014).
[9] V. Oppen, J. Madeleine, L. M. Peplow, S. Kininmonth, and R. Berkelmans, Molecular Ecology 20, 4899 (2011).
[10] D. Brockmann and D. Helbing, Science 342, 1337 (2013).
[11] M. D. Conover, C. Davis, E. Ferrara, K. McKelvie, F. Menczer, and A. Flammini, PloS one 8, e55957 (2013).
[12] F. Pozzi, T. Di Matteo, and T. Aste, Scientific reports 3, 1665 (2013).
[13] J. M. Hughes, N. J. Foti, D. C. Krakauer, and D. N. Rockmore, Proceedings of the National Academy of Sciences 109, 7682 (2012).
[14] D. Shalymov, O. Granichin, L. Klebanov, and Z. Volkovich, Expert Systems with Applications 61, 145 (2016).
[15] S. Fortunato, *Community detection in graphs* (2010), 0906.0612.
[16] W. M. Song, T. Di Matteo, and T. Aste, PLoS ONE 7 (2012), ISSN 19326203, 1110.4477.
[17] R. N. Mantegna, European Physical Journal B 11, 193 (1999), ISSN 1434-6028, 9802256v1.
[18] M. Tumminello, T. Aste, T. Di Matteo, and R. N. Mantegna, Proceedings of the National Academy of Sciences 102, 10421 (2005), ISSN 0027-8424, http://www.pnas.org/content/102/30/10421.full.pdf.
[19] G. P. Massara, T. Di Matteo, and T. Aste, Journal of Complex Networks 5, 161 (2016), ISSN 20511329, 1505.02445.
[20] I. J. Ramasco and B. Gonçalves, Physical Review E - Statistical, Nonlinear, and Soft Matter Physics 80 (2009), ISSN 15393755, 0609776.
[21] J. Shalymov, O. Granichin, L. Klebanov, and Z. Volkovich, Expert Systems with Applications 61, 145 (2016).
[22] S. Fortunato, *Community detection in graphs* (2010), 0906.0612.
[28] F. Radicchi, J. J. Ramasco, and S. Fortunato, Physical Review E - Statistical, Nonlinear, and Soft Matter Physics 83 (2011), ISSN 15393755, 1009.2913.
[29] P. B. Slater, Proceedings of the National Academy of Sciences 106, E66 (2009), ISSN 0027-8424, 0904.2389.
[30] M. Tumminello, S. Micciciè, F. Lillo, J. Piilo, and R. N. Mantegna, PLoS ONE 6 (2011), ISSN 19326203, 1008.1414.
[31] M. X. Li, V. Palchykov, Z. Q. Jiang, K. Kaski, J. Kertész, S. Micciciè, M. Tumminello, W. X. Zhou, and R. N Mantegna, New Journal of Physics 16 (2014), ISSN 13672630, 1403.3785.
[32] G. Iori and R. N. Mantegna (2018).
[33] R. Miller, Simultaneous Statistical Inference (Springer Verlag New York, 1981), 2nd ed., ISBN 0-387-90548-0.
[34] J. Zhao, L. Miao, J. Yang, H. Fang, Q.-M. Zhang, M. Nie, P. Holme, and T. Zhou, Scientific reports 5, 12261 (2015).
[35] A. Barrat, M. Barthélemy, and A. Vespignani, Physical review letters 92, 228701 (2004).
[36] J. Haigh, Polya Urn Models, vol. 172 (2009), ISBN 9781420059830.
[37] Y. Benjamini and Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing (1995), 95/57289.
[38] T. Perneger, BMJ 316, 1236 (1998).
[39] W. Noble, Nature Biotechnology 27, 1135 (2009).
[40] Us airports network dataset – KONECT, [http://konect.uni-koblenz.de/networks/opsahl-usairport](http://konect.uni-koblenz.de/networks/opsahl-usairport) (2017).
[41] T. Opsahl, Why anchorage is not (that) important: Binary ties and sample selection (2011), URL [http://wp.me/poFcY-Vw](http://wp.me/poFcY-Vw).
[42] Data available at [http://www.wiod.org/home](http://www.wiod.org/home) (2016).
[43] C. Bachmaier, U. Brandes, and B. Schleiper, in International Symposium on Algorithms and Computation (Springer, 2005), pp. 1110–1121.
[44] D. Blackwell and J. B. MacQueen, The Annals of Statistics 1, 353 (1973), ISSN 0090-5364, arXiv:1306.3979v1.
[45] D. Grady, C. Thiemann, and D. Brockmann, Nature Communications 3 (2012), ISSN 20411723, 1110.3864.
[46] E. Dietzenbacher, B. Los, R. Stehrer, M. Timmer, and G. de Vries, Economic Systems Research 25, 71 (2013), ISSN 09535314.
[47] F. Cerina, Z. Zhu, A. Chessa, and M. Riccaboni, PLoS ONE 10 (2015), ISSN 19326203, 1407.0225.
[48] W. Li, D. Y. Kenett, K. Yamasaki, H. E. Stanley, and S. Havlin, arXiv preprint arXiv:1408.0443 (2014).
[49] F. Portella-Carbó, Economic Systems Research 28, 95 (2016), ISSN 14695758.
[50] Data available at [https://www.bts.gov](https://www.bts.gov) (2017).
[51] In the case of a link between a node with degree $k = 1$ and a node with $k > 1$ we keep the link only if $\pi P < \alpha_n/L$ for the node with degree greater than one.