RESEARCH

Learning Attribute and Homophily Measures Through Random Walks

Nelson Antunes[1*], Sayan Banerjee[2], Shankar Bhamidi[2] and Vladas Pipiras[2]

[*]Correspondence:
nantunes@ualg.pt
[1]Center for Computational and Stochastic Mathematics, University of Lisbon, Avenida Rovisco Pais, 1049-001, Lisbon, Portugal
Full list of author information is available at the end of the article

Abstract

We investigate the statistical learning of nodal attribute functionals in homophily networks using random walks. Attributes can be discrete or continuous. A generalization of various existing canonical models, based on preferential attachment is studied (model class $\mathscr{P}$), where new nodes form connections dependent on both their attribute values and popularity as measured by degree. An associated model class $\mathscr{U}$ is described, which is amenable to theoretical analysis and gives access to asymptotics of a host of functionals of interest. Settings where asymptotics for model class $\mathscr{U}$ transfer over to model class $\mathscr{P}$ through the phenomenon of resolvability are analyzed. For the statistical learning, we consider several canonical attribute agnostic sampling schemes such as Metropolis-Hasting random walk, versions of node2vec (Grover and Leskovec, 2016) that incorporate both classical random walk and non-backtracking propensities and propose new variants which use attribute information in addition to topological information to explore the network. Estimators for learning the attribute distribution, degree distribution for an attribute type and homophily measures are proposed. The performance of such statistical learning framework is studied on both synthetic networks (model class $\mathscr{P}$) and real world systems, and its dependence on the network topology, degree of homophily or absence thereof, (un)balanced attributes, is assessed.

Keywords: attributed networks; homophily; network model; resolvability; random walk samplings; discrete and continuous attributes; learning attribute functionals.

Introduction

Attributed networks, namely graphs in which nodes and/or edges have attributes, are at the center of network-valued datasets in many modern applications. For example, in real-world network datasets most nodes have values of characteristics of interest; in social networks, users have attributes such as "gender", "age", "language"; in citation networks, articles are classified by the main subject, field, subfield, keywords. Networks also differ in the range of attributes values (cardinality), their types (discrete or continuous) and the size of each group. In one direction, machine learning pipelines such as network representation learning [1], clustering [2], classification [3], and community detection [4] have been developed to study the entire network. Another recent direction, specifically related to attributed network valued data, is the use of attribute information, in addition to graph topological information, in improving the performance of exploratory data analytic techniques such as community detection [5] or link prediction tasks [6]. Both papers, through careful development of methodological analysis using graph regularization and non-negative matrix factorization, and through detailed empirical analysis, show sig-

nificant improvement for such machine learning pipelines via incorporating node attribute information. Driven by the scale of data, the main motivation of this paper is network sampling, where limited explorations based on random walks are used to learn network level functionals of attributes.

In real-world networks, the attributes of a node will co-vary and are not independent. One standard phenomenon in many such real world systems is *homophily* [7, 8, 9], i.e., node pairs with similar attributes being more likely to be connected than node pairs with discordant attributes. For instance, many social networks show this property, which is the tendency of individuals to associate with others who are similar to them; e.g., with respect to the gender, ethnicity, political ideologies. Furthermore, the distribution of user attributes over the network is usually uneven, with coexisting groups of different sizes, e.g., one ethnic group may dominate others [10]. On the other hand, another co-variation across neighbors is due to heterophily, where nodes with the same attribute type value repel each other.

Performance of network sampling algorithms in such settings has received some attention including: the bias of several sampling methods in conserving position of nodes and visibility of groups [11]; the effect of homophily on centrality measures and visibility of minority groups and fairness questions [12]. More recently the synthetic models that motivate this paper were used in [13] to understand the inequality of node ranking algorithms (e.g. as measured by the Gini coefficient) as well as inequity (e.g. by contrasting the percentage of a given attribute amongst the most popular k%-age of nodes with the true demographic percentage of that group), in particular trying to understand the foundational characteristics of network evolution such as homophily or preferential attachment in (quoting [13]) "reducing, replicating or amplifying" representation of specific groups by these ranking algorithms. In a different direction, [10] uses these synthetic models to understand the accuracy of semi-supervised machine learning tasks such as learning/prediction of attribute labels given partial information on the labels of a subset of seeded vertices; the goal is to understand the impact of homophily/heterophily and preferential attachment driven growth characteristics of the underlying network on the accuracy of a host of popular relational classifiers and collective inference algorithms.

This paper is motived by the lack of theoretical results in the analysis of attribute network models with homophily and the development of a learning framework to estimate attribute functionals in real networks. We investigate the following research questions (RQ).

RQ1: How to analyze and extend the existing network models with homophily and derive the main functionals of interest?

We describe a generalization of the directed preferential attachment model with homophily (called model class $\mathscr{P}$) formulated in [12] where new nodes connect to existing ones based on the attributes of both end points of the potential edge and centrality of the existing vertex. The network model can generate scale-free networks with discrete or continuous attributed nodes, and different intensities of homophily. The dynamics of the network is the following. Starting from a fully connected cluster of nodes with attributes, each node that arrives has attribute generated independently according to a given distribution and connects to a fixed (constant) number of nodes. The probability that a new node connects to an existing

node is proportional to the product of the degree (to the power of a parameter) with a function that measures the propensity of the two nodes attributes to interact. Thus, the model encodes the interplay between the two main mechanisms of tie formation found in social networks: preferential attachment and homophily. Given the importance of this model in applications, theoretical analysis of this model including stability properties of heterophily and homophily statistics are of great importance; yet till date the only functional amenable to theoretical analysis has been degree distribution [12, 14]. We describe a related model of network evolution (called model class $\mathscr{U}$) which is much more amenable to theoretical analysis and a phenomenon we term *resolvability* which enables one to transfer results from model class $\mathscr{U}$ to model class $\mathscr{P}$; in this paper we specialize to large network limits for degree distribution for an attribute type and homophily and heterophily statistics, deferring a full treatment to [15].

RQ2: How to use the existing link trace algorithms to sample the network and take into account the attributes of nodes?

Uniform random sampling of nodes or edges is the "gold standard", providing unbiased estimates of corresponding attribute functionals. However, owing to both computational and privacy issues in social networks and other settings, such sampling is often infeasible. Other networks that allow random access limit the rate of API (Application Program Interface) calls implying that creating a sample of sufficient size takes a prohibitive time. In these cases, link trace sampling, such as random walks (RWs) are typically used; see references in [16, 17] for estimation of functionals such as degree distribution and clustering. However, much less is known in the context of estimating quantities influenced by attribute types in homophily networks.

In this work, we consider several existing canonical attribute agnostic sampling schemes proposed in the literature (that do not use the attribute type of nodes to construct the sample) such as Metropolis-Hasting random walk and versions of node2vec [18] that incorporate both classical random walk and walks with non-backtracking propensities. These random walks have been designed to preserve structural properties of the network in the sample, such as high degree nodes, clustering, diameter and not the different types of node attributes. We are interested not only in estimating the proportion of nodes with a given attribute but also in the structural properties of the sub-network spanned by vertices of a specified attribute type including the degree distribution and homophily measures. Our main contribution here is to show that random walks that use edge weights can be attribute aware samplers through the proposal of variants of node2vec where edge weights depend on attributes of its end nodes. This will be especially useful in homophilic networks for analyzing geometric properties involving nodes with minority attributes.

RQ3: How to estimate the attribute functionals and homophily measures through the sampling schemes and evaluate their performance?

We propose estimators for attribute functionals and homophily measures that are based on correcting the bias of the empirical sample quantities through the use of stationary distribution of the RWs associated in sampling nodes and edges.

We study the performance of the considered random walk sampling schemes in terms of estimation error of the attribute distributions and homophily measures

across the following four dimensions in both synthetic networks using the model class $\mathscr{P}$ and real world settings: **(a)** Inherent homophilic propensity of the network and underlying density of attributes; **(b)** Impact of centrality of nodes as measured by degree in the evolution of the network; **(c)** Nonlinear impact of incorporating "escape echo chamber" mechanisms in random walks by encouraging walks to jump across edges with discordant attributes; **(d)** Impact of reducing the backtracking propensity to encourage walks to explore more of the network. We find that (*i*) RWs with attribute dependent weights can perform better over attribute agnostic RWs in homophilic networks; (*ii*) the weights need to balance the movements between/within nodes with different/same attributes; (*iii*) non-backtracking improves performance, especially in conjunction with attribute dependent weights and low edge density; (*iv*) methods seem to work comparably well for synthetic and real networks.

This paper is a significant extension of the conference paper [19] including: (a) appreciable expansion of the theoretical developments to the network models described in [19], including describing the notion of *resolvability* of such models which allows one to connect them to a different class of models for which asymptotic analysis for a wide range of functionals, such as degree exponent for an attribute type, homophily and heterophily statistics can be undertaken; (b) substantial expansion of the methodological development of the paper, including a new class of functionals (degree distribution for an attribute and homophily measures) to be estimated through network sampling schemes; (c) new network sampling schemes from node2vec variants; (d) further applications of the methodology developed to new network data for evaluation and comparison; and (e) a final section with extensions and future directions of the work.

Attributed Network Models and Homophily Functionals

As described above, synthetic models have been used to great effect in understanding the structure and evolution of attributed networks and the impact of ranking, sampling and classification algorithms in such settings. The overarching goal in this section is to describe an extension of the canonical (linear) attributed network models currently considered in the literature. We refer the interested reader to [12, 13, 10] and the references therein for further discussion on motivations and use of such models. More concretely in this section:

(a) We will describe the main synthetic model, termed non-linear preferential attachment (NLPA) model with homophily, and referred to for the rest of the paper as model class $\mathscr{P}$.

(b) We will give concrete formulations of key network functionals measuring homophily between different groups.

(c) Understanding (large network) asymptotics for model class $\mathscr{P}$ is non-trivial. We will introduce a related model (referred to as model class $\mathscr{U}$), that seems significantly more amenable to analysis, formalize a notion called *resolvability*, connecting model classes $\mathscr{P}$ and $\mathscr{U}$ and then describe the explicit results that can be derived for model class $\mathscr{P}$, at least in the linear case using $\mathscr{U}$. Technical justifications of these connections can be found in [15].

Table 1 Summary of the main notation.

Notation	Description
$\mathscr{P}(\alpha,\mu,f)$	model class $\mathscr{P}$ (non-linear preferential attachment model with homophily)
$\mathcal{A}$	attribute space
μ	attribute distribution of an arriving node
α	preferential attachment parameter
$f(a,a')$	propensity of a pair of nodes with attributes a and a' to interact
m	number of edges a new node entering the system connects to pre-existing nodes
$\deg_i(v,n)$	degree of v at time n when i of the edges of v_{n+1} have connected to the network
$\mathscr{U}(\alpha,\nu,f)$	model class $\mathscr{U}$
ν	resolvable measure
$\mathcal{E}$	set of edges of the network
N	number of nodes in the network
$\mathcal{V}_a$	set of nodes of type a
$\mathcal{E}_{aa'}$	set of edges between nodes of attributes a and a'
D_a	dyadicity of nodes with attribute a
$H_{aa'}$	heterophilicity between nodes with attributes a and a'
$p(a)$	proportion of nodes with attribute a
$\lvert.\rvert$	number of elements of a set
$p(k\lvert a)$	proportion of nodes of degree k having attribute a
$\widehat{.}$	estimator of a quantity
d_i	degree of node i
π_i	probability of sampling node i
$\pi_{(i,j)}$	probability of sampling edge (i,j)
w_{ij}	weight of edge (i,j)
θ	propensity of N2V to backtrack
γ (β)	propensity of a N2V to reach a (non-)common neighbor of the currently visited node and the previously visited node
δ	spectral gap

Fix an attribute (or latent) space $\mathcal{A}$ with probability measure μ. Fix a (potentially asymmetric) function $f : \mathcal{A} \times \mathcal{A} \to \mathbb{R}_+$ which measures propensities of node pairs to interact based on their attributes. Fix $\alpha \geq 0$ describing the role of degree in measuring popularity and integer $m \geq 1$ denoting the number of edges a new vertex has when entering the system, to connect to pre-existing vertices. In principle m could be random and/or dependent on the attribute type, but for simplicity and to match existing literature (e.g. [12]) we focus on the fixed m setting (see [15] for results when m is attribute dependent). Let N be the number of nodes (vertices) in the network. In the model class $\mathscr{P}$, nodes $\{v_n : 1 \leq n \leq N\}$ enter the system sequentially starting at $n = 1$ with a base connected graph $\mathcal{G}_1$ (with every node having an attribute in $\mathcal{A}$) with dynamics:

(i) Every node v_n has attribute $a(v_n) \in \mathcal{A}$ generated independently using μ.

(ii) Node v_n enters the system with m edges.

(iii) The dynamics for connecting each of the m edges are recursively defined as follows: suppose the network has been constructed till stage n with structure $\mathcal{G}_n$. For any n and $0 \leq i \leq m-1$ and $v \in \mathcal{G}_n$, let $\deg_i(v,n)$ denote the degree of v at time n when i of the edges of v_{n+1} have connected to $\mathcal{G}_n$. Conditional on $\mathcal{G}_n$ and stage i, the probability that the $(i+1)$th edge of v_{n+1} connects to $v \in \mathcal{G}_n$ is proportional to:

$$P_{v_{n+1}v} \propto f(a(v), a(v_{n+1}))[\deg_i(v,n)]^{\alpha}. \tag{1}$$

Once this edge has connected, all the degrees are updated and the above dynamics is repeated till all m edges have connected to $\mathcal{G}_n$. When $m = 1$, then each new vertex has only one edge to connect to the network and in this case we write $\deg(v,n) := \deg_0(v,n)$.

We will refer to this as model class $\mathscr{P}$ (or $\mathscr{P}(\alpha, \mu, f)$ when we want to specify all the parameters; we suppress dependence on m to ease notation) and sometimes write $\{\mathcal{G}_n : 1 \leq n \leq N\} \sim \mathscr{P}(\alpha, \mu, f)$. The model (1) extends various existing models including: Barabási-Albert model [20] ($f \equiv 1$, $\alpha = 1$), sublinear PA [21] ($f \equiv 1$, $0 < \alpha < 1$), PA with multiplicative fitness [22] ($f(a, a') = a$, $\alpha = 1$), scale free homophilic model [23] ($f(a, a') = 1 - |a - a'|$, $\mathcal{A} = [0, 1]$, $\alpha = 1$), and geometric versions with $\alpha = 1$, a compact metric space $\mathcal{A}$ and an appropriate function f of the distance [24, 14]. Most existing studies focus on asymptotics for either the degree distribution or maximal degree.

Homophily functionals

When the latent space $\mathcal{A} = \{1, 2, \ldots, K\}$ is finite, one can define macroscopic measures of homophily, and conversely heterophily [25], from an observed network $\mathcal{G}$ (either synthetic or empirically observed) on N nodes as follows. Let $\mathcal{E}$ denote the total edge set; for $a \in \mathcal{A}$, let $\mathcal{V}_a$ be the set of nodes of type a, and for $a, a' \in \mathcal{A}$, let $\mathcal{E}_{aa'}$ be the set of edges between nodes of types a and a'. Let $p = |\mathcal{E}|/\binom{N}{2}$ be the edge density. For $a \in \mathcal{A}$, dyadicity

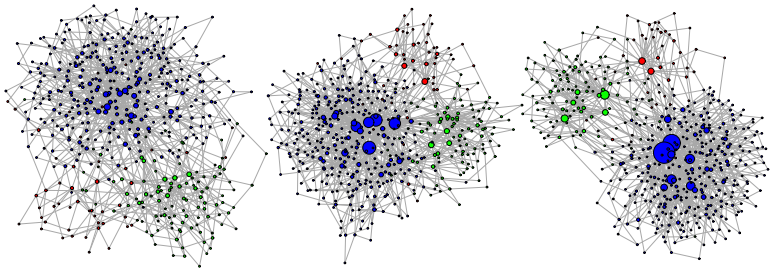
$$D_a = |\mathcal{E}_{aa}| \Big/ \left(\binom{|\mathcal{V}_a|}{2} p \right) \tag{2}$$

measures the contrast in edges within the cluster of nodes a as compared to a setting where all edges are randomly distributed; thus $D_a > 1$ signals homophilic characteristics of type a nodes while $D_a < 1$ signifies heterophilic nature of type a nodes. Similarly, for $a \neq a'$, heterophilicity

$$H_{aa'} = |\mathcal{E}_{aa'}|/(|\mathcal{V}_a||\mathcal{V}_{a'}|p) \tag{3}$$

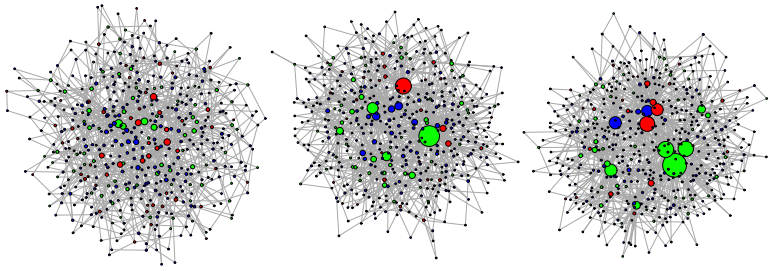
denotes propensity of type a nodes to connect to type a' nodes as contrasted with random placement of edges with probability equal to the global edge density. If $H_{aa'} < 1$, nodes of opposite labels do not tend to be connected (homophilic); if $H_{aa'} > 1$, there are more connections between nodes of different labels a and a' (heterophilic).

Illustrations of homophilic synthetic networks of the model class $\mathscr{P}(\alpha, \mu, f)$ generated from (1) are given in Fig. 1. The total number of nodes is $N = 1000$ and each node has an attribute in $\mathcal{A} = \{1, 2, 3\}$ according to the probability mass function (p.m.f.) $\mu = (0.7, 0.2, 0.1)$; the propensities of node pairs to connect based on their attributes are $f(a, a) = 0.8$, $f(a, a') = 0.1$, $a \neq a' = 1, 2, 3$ and $m = 2$. The networks are plotted for different values of α in Fig. 1(a)–1(c). For instance, with $\alpha = 0.2$, the corresponding homophily measures are $D_1 = 1.364$, $D_2 = 3.038$, $D_3 = 7.38$, $H_{12} = 0.336$, $H_{13} = 0.386$ and $H_{23} = 0.399$. Figure 2 shows the case of heterophilic networks with $N = 1000$ of the model class $\mathscr{P}(\alpha, (0.7, 0.2, 0.1), f)$ for different values of α with $f(a, a) = 0.2$, $f(a, a') = 0.4$, $a \neq a' = 1, 2, 3$ and $m = 2$. For $\alpha = 1$, the homophily measures are $D_1 = 0.750$, $D_2 = 0.772$, $D_3 = 0.750$, $H_{12} = 1.479$, $H_{13} = 1.615$ and $H_{23} = 1.873$.



(a) (b) (c)

Figure 1 Networks generated by the model class $\mathscr{P}(\alpha, \mu, f)$ with homophily where $N = 1000$, $\mu = (0.7, 0.2, 0.1)$, $f(a, a) = 0.8$, $f(a, a') = 0.1$, $a \neq a' = 1, 2, 3$ and $m = 2$ (a) $\alpha = 0.2$, (b) $\alpha = 1$, (c) $\alpha = 1.2$.



(a) (b) (c)

Figure 2 Networks generated by the model class $\mathscr{P}(\alpha, \mu, f)$ with heterophily where $N = 1000$, $\mu = (0.7, 0.2, 0.1)$, $f(a, a) = 0.2$, $f(a, a') = 0.4$, $a \neq a' = 1, 2, 3$ and $m = 2$ (a) $\alpha = 0.2$, (b) $\alpha = 1$, (c) $\alpha = 1.2$.

Model class $\mathscr{U}$ and rationale

While model class $\mathscr{P}$ has been heavily used in applications, deriving large network asymptotics of functionals is non-trivial. Next we will describe a related network model (model class $\mathscr{U}$), the rationale for why this might be more amenable to analysis, and then formalize situations where given $\mathscr{P}$, one can construct (using as input the parameters α, μ, f from $\mathscr{P}$), a corresponding model in class $\mathscr{U}$ such that properties of $\mathscr{P}$ can be read off from (the more easily analyzable) $\mathscr{U}$. For most of this discussion we will only consider the $m = 1$ setting, albeit the formulae for asymptotics for various functionals considered below seem to extend, at least in simulations, in a straightforward manner to general m setting.

Since the general setting (with "continuous" attribute space) is more technical, let us explain the basic rationale in the simpler discrete setting where $\mathcal{S} = [K] := \{1, 2, \ldots, K\}$ so that μ is a p.m.f.. Fix a (potentially and in most cases different from μ) p.m.f. ν and consider the attributed network model $\left\{\tilde{\mathcal{G}}_n : n \geq 0\right\}$ with dynamics:

$$\mathbb{P}\left(a(v_{n+1}) = a^\star, v_{n+1} \rightsquigarrow v | \tilde{\mathcal{G}}_n\right) := \frac{\nu(a^\star) f(a(v), a^\star)[\deg(v, n)]^\alpha}{\sum_{a \in [K]} \sum_{v' \in \tilde{\mathcal{G}}_n} \nu(a) f(a(v'), a)[\deg(v', n)]^\alpha}. \quad (4)$$

Note that the above model is invariant to scaling in ν, so it will be convenient to allow ν to be a general weight sequence instead of normalizing it to be a probability measure.

The above belongs to a general class of models defined below that we will refer to as $\mathscr{U}(\gamma, \nu, f)$. Thus, here the p.m.f. ν plays the role of a weight and further, unlike the model $\mathscr{P}$ where each new arriving vertex has attribute sampled independently from the current state of the network, here the distribution of new vertices is closely dependent on the entire state of the current network.

Rationale for technical tractability: Tabling the issue of connection with $\mathscr{P}$ for the next sections, first note that $\mathscr{U}$ can be simulated via dynamics where every vertex essentially **behaves independently** ((c) below). In brief, if one wanted to simulate model class $\mathscr{U}$ starting from one vertex of type a, then this can be done as follows:

(a) Every vertex v that enters the system (starting with the root of type a) gives birth independently to child nodes with attributes in continuous time, connected to the vertex.
(b) For a node of type a, conditional on its degree d, the rate of reproduction of a child node of type a' is $\nu(a)f(a, a')d^{\alpha}$.
(c) Reproduction dynamics is independent across nodes.

Write $\{\mathrm{BP}(t) : t \geq 0\}$ for the (continuous time) process and for any $n \geq 1$, T_n be the (random) time such that the size $|\mathrm{BP}(T_n)| = n$. (BP stands for Branching Process.) Then it is easy to check that $\{\mathrm{BP}(T_n) : 1 \leq n \leq N\}$ has the same distribution as $\left\{\tilde{\mathcal{G}}_n : 1 \leq n \leq N\right\} \sim \mathscr{U}(\alpha, \nu, f)$. Further the independence in the evolution makes this model much more amenable to analysis, yielding asymptotic information for the process BP and thus the model $\mathscr{U}$.

Resolvability

Note that the main model of interest, both as a synthetic test bed in this paper, and in pre-existing work, is the model class $\mathscr{P}$. The main goal of this section is to formalize a connection between model classes $\mathscr{P}$ and $\mathscr{U}$. Given $\left\{\tilde{\mathcal{G}}_n : 0 \leq n \leq N\right\} \sim \mathscr{U}(\alpha, \nu, f)$, for $n \geq 1$ define $\tilde{\pi}_n = \sum_{t=1}^{n} \delta\{a(v_t)\}$, i.e. the empirical measure of attributes in $\tilde{\mathcal{G}}_n$.

Now say that model $\mathscr{P}(\alpha, \mu, f)$ is *resolvable* if there exists ν such that for the model class $\mathscr{U}(\alpha, \nu, f)$, the empirical measures of attribute types satisfy: $\tilde{\pi}_n \to \mu$ as $n \to \infty$. In words, one can chose a weight measure ν such that the corresponding dynamics for $\mathscr{U}$ with the same α and f drives the empirical distribution to the limiting empirical distribution μ of model class $\mathscr{P}$ (since every new vertex has attribute distribution μ independent of the network evolution).

Resolvability in the linear finite attribute case

The linear case ($\alpha = 1$) with a finite attributes $\mathcal{S} = [K]$ turns out to be completely resolvable under the following.

Assumption: Assume the sampling measure $\mu = (\mu_1, \ldots, \mu_K)$ has all entries strictly positive and assume the affinity kernel $f(a, a') > 0$, $\forall a, a' \in [K]$.

Fix a model class $\mathscr{P}(\alpha = 1, \mu, f)$ satisfying the above Assumption. Let $\mathcal{P}([K])$ denote the $K - 1$ dimensional simplex of probability mass functions on $[K]$. Define

(in the interior of $\mathcal{P}([K])$) the function:

$$V_\mu(y) := 1 - \frac{1}{2}\sum_{j\in\mathcal{S}} \mu_j \left(\log(y_j) + \log(\sum_{k\in\mathcal{P}} y_k f(k,j)) \right), \qquad y \in \mathcal{P}([K]).$$

By [14, P8], under the above Assumption, $V_\mu(\cdot)$ has a *unique* minimizer $\eta := \eta(\mu) = (\eta_1(\mu), \ldots, \eta_K(\mu))$ in the interior of $\mathcal{P}(\mathcal{S})$. Now, define

$$\nu_a := \frac{\mu_a}{\sum_{l=1}^K f(l,a)\eta_l}, \qquad \phi_{a,b} := f(a,b)\nu_b, \qquad \phi_a := \sum_{b=1}^K \phi_{a,b} = 2 - \frac{\mu_a}{\eta_a}, \quad (5)$$

where the final identity follows from [14, P8]. Let $\nu = (\nu_1, \ldots, \nu_K)$. Then the following paraphrases some of the results in [15]:

(a) Under the above Assumption, model $\mathscr{P}(\alpha = 1, \mu, f)$ is resolvable with one resolving measure ν given as above. This implies, in particular, local functionals (such as degree distribution PageRank) converge to the same limits as those for $\mathscr{U}(\alpha = 1, \nu, f)$. Two specific implications are given next.

(b) For each $a \in [K]$, the empirical p.m.f. of vertice degrees of type $\mathbf{p}_n^a \xrightarrow{P} \mathbf{p}_\infty^a$ where the limit p.m.f. has tail exponent $\mathbf{p}_\infty^a(k) \sim k^{1+2/\phi_a}$ as $k \to \infty$.

(c) Using the objects defined in (5), define the matrix

$$\mathbf{M} = \left(\mathbf{M}_{a,b} := \frac{\phi_{a,b}}{2 - \phi_a} \right)_{a,b\in[K]}. \quad (6)$$

Then the homophily and heterophily statistics $\{D_{n,a} : a \in [K]\}$ and $\left\{H_{n,(a,a')} : a \neq a' \in [K]\right\}$ satisfy the asymptotics,

$$D_{n,a} \xrightarrow{P} \frac{[\mathbf{M}]_{a,a}}{\mu_a}, \qquad H_{n,(a,a')} \xrightarrow{P} \frac{1}{2}\left[\frac{[\mathbf{M}]_{a',a}}{\mu_a} + \frac{[\mathbf{M}]_{a,a'}}{\mu_{a'}} \right] \quad (7)$$

Remark 1 Result (b) above was previously derived in [14] using stochastic approximation techniques.

The results above are illustrated numerically in Fig. 3 and Tables 2 and 3. We fixed the model class $\mathscr{P}(1, (0.7, 0.2, 0.1), f)$, where $f(a,a) = 0.8$, $f(a,a') = 0.1$, for $a \neq a' = 1, 2, 3$ and $m = 1$. The model is resolvable with resolving measure ν approximately equal to $(0.742, 0.189, 0.069)$. We generate the model classes $\mathscr{P}$ and $\mathscr{U} = (\alpha, \nu, f)$ using (1) and (4), respectively, for different network sizes. Fig. 3 shows the degree distributions of attribute 2 for both models which are getting closer as N increases. In the limit they converge to the same p.m.f.. We fit a power-law distribution function using a maximum likelihood approach to the empirical degree distribution tail per attribute of the model class $\mathscr{P}$ for each network size. The respective tail exponents are shown in Table 2 with the asymptotic limit p.m.f. tail exponent $1 + 2/\phi_a$. Finally, the empirical and asymptotic dyadicity and heterophilicity measures, respectively, (2), (3) and (7), are given in Table 3. The results show that complicated functionals of the model class $\mathscr{P}$ can be easily approximated with good precision even for moderate network sizes.

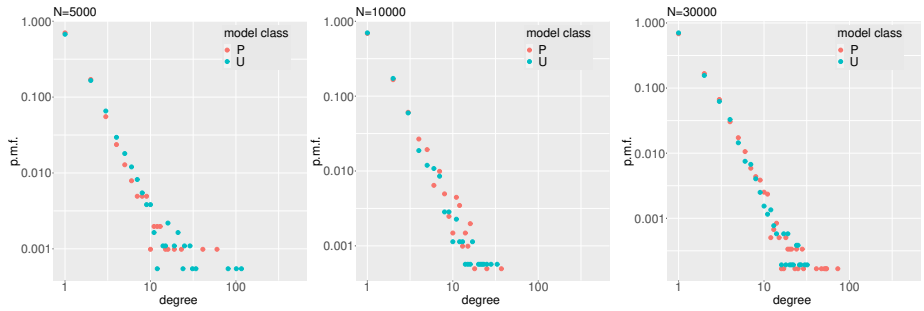


Figure 3 Degree distribution of attribute 2 under model classes $\mathscr{P}(\alpha, \mu, f)$ and $\mathscr{U}(\alpha, \nu, f)$ for different network sizes N, where $\alpha = 1$, $\mu = (0.7, 0.2, 0.1)$, $\nu \approx (0.742, 0.189, 0.069)$, $f(a, a) = 0.8$, $f(a, a') = 0.1$, $a \neq a' = 1, 2, 3$ and $m = 1$.

Table 2 Tail exponent of the degree distribution per attribute of model class $\mathscr{P}(\alpha, \mu, f)$ for different network sizes N and asymptotically ($N \to \infty$), where $\alpha = 1$, $\mu = (0.7, 0.2, 0.1)$, $f(a, a) = 0.8$, $f(a, a') = 0.1$, $a \neq a' = 1, 2, 3$ and $m = 1$.

Attribute	1	2	3
Asym. model class $\mathscr{P}$	2.892	3.256	3.782
$N = 5000$	2.933	3.352	3.547
$N = 10000$	2.950	3.230	4.079
$N = 30000$	2.983	3.310	3.742

Table 3 Homophily measures of model class $\mathscr{P}$ for different network sizes N and asymptotically ($N \to \infty$), where $\alpha = 1$, $\mu = (0.7, 0.2, 0.1)$, $f(a, a) = 0.8$, $f(a, a') = 0.1$, $a \neq a' = 1, 2, 3$ and $m = 1$.

	D_1	D_2	D_3	H_{12}	H_{13}	H_{23}
Asym. model class $\mathscr{P}$	1.369	3.183	4.038	0.3074	0.4059	0.4633
$N = 5000$	1.391	3.115	3.961	0.312	0.393	0.4362
$N = 10000$	1.384	3.226	3.459	0.291	0.433	0.448
$N = 30000$	1.363	3.185	3.774	0.316	0.415	0.476

Random Walk Samplings in Attributed Networks

Since many real-world networks can only be crawled, in the sense that only the neighbors of the current visited node can be explored, we consider sampling procedures that are based on random walks. They are also a core technique for constructing various algorithms to extract information on networks, such as community detection, ranking of nodes and edges, and dimension reduction. We introduce well-known random walks which are attribute agnostic. These random walks have been designed to preserve structural properties of the network and not the representativeness of node attributes in the sample. We are interested (see next section) in estimating the attribute distribution but also structural properties (node degrees) depending on the node attributes. We show next that some random walks that use edge weights can be attribute aware samplers. This will be especially useful in homophilic networks. Throughout this section, for graph $\mathcal{G}$ and node $i \in \mathcal{G}$, d_i will denote its degree. We assume a static graph and that only limited set of initial seed nodes $i \in \mathcal{G}$ that initializes the random walk are available. When we say that a node is sampled, it means that its attribute $a(i)$ (and degree $d(i)$ dependent on the quantities of interest) is added to the sample.

Metropolis Hastings Random Walk (MHRW). At each step, if the walk is currently at node i, a neighbor j is selected uniformly at random and the proposed

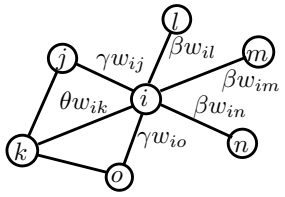


Figure 4 Node2vec node transitions. The random walk has transitioned from k to i and is now evaluating the next step out of i.

move to j is accepted with probability $\min(1, d_i/d_j)$, else the walk stays at i. Thus proposed moves towards a node of smaller degree are always accepted whilst we reject some of the proposed moves towards higher degree nodes. It is easy to check that the stationary distribution is uniform over the node set, i.e.,

$$\pi_i = 1/N, \quad 1 \leq i \leq N. \tag{8}$$

The stationary distribution over the edge set is

$$\pi_{ij} = \frac{1}{Nd_i}, \quad (i,j) \in \mathcal{E}. \tag{9}$$

Node2vec (N2V). As proposed in [18], in full generality, the transitions of N2V depend on the neighborhood both of the currently visited node, and the node visited prior to the current node. Let the previously and currently visited nodes be k and i, resp. The next visited node j is chosen according to the transition probability proportional to:

$$p(j|k,i) \propto \begin{cases} \beta w_{ij}, & k \neq j, (k,j) \notin \mathcal{E}, \\ \gamma w_{ij}, & k \neq j, (k,j) \in \mathcal{E}, \\ \theta w_{ij}, & k = j, \end{cases}$$

where w_{ij} is the weight of edge (i,j), θ is the parameter that represents the propensity for the random walk to backtrack, γ is the quantifying probability of reaching a common neighbor of the currently visited node and the node visited in the last step, and β is the parameter of exploring any of other neighbor – see Fig. 4. N2V is a second order Markov chain. We now describe specific variants of this random walk which includes some classical versions.

Node2vec-1 (N2V-1): If the network is undirected, unweighted and $\theta = \beta = \gamma$, one obtains the classical RW with the well-known stationary distributions,

$$\pi_i = \frac{d_i}{2|\mathcal{E}|}, \qquad \pi_{ij} = \frac{1}{|\mathcal{E}|}. \tag{10}$$

Node2vec-2 (N2V-2): If the network is undirected and $\theta = \beta = \gamma$, one obtains a weighted RW. This walk can use node attributes through weights in contrast to N2V-1. We assume that for each sampled node i, we have access to the attributes

of the neighbors of i. If there is a connection between i and j, the weight w_{ij} is a function of $a(i)$ and $a(j)$. In a homophilic network, setting w_{ij} to a lower value if nodes have equal attributes encourages the sampling of nodes with different attributes. The stationary distributions in this case are given by

$$\pi_i = \frac{\sum_j w_{ij}}{\sum_k \sum_j w_{kj}}, \qquad \pi_{ij} = \frac{w_{ij}}{\sum\sum_{k<l} w_{kl}}. \tag{11}$$

Node2vec-3 (N2V-3): If the network is undirected, without self-loops, multiple edges and $\beta = \gamma$, $\theta > 0$, with equal weights w_{ij}, the stationary distributions for nodes and edges are given by (10) [26]. With small θ, the walk approaches the non-backtracking random walk avoiding 2-hop redundancy in the sample.

Node2vec-4 (N2V-4): We consider next the combination of the last two schemes, with $\beta = \gamma$, $\theta > 0$ and weights w_{ij} dependent on the attributes of i and j. In this setting, one major technical hurdle is that, unlike the settings above, there is no explicit formula for the stationary distributions. Analogous to the stationary distributions for N2V-3 matching the usual RW in the stationary regime, it is expected that especially in the small θ setting, the stationary distributions can still be approximated by those in (11). We explore the efficacy of these approximations for moderate size synthetic networks below.

Node2vec-5 (N2V-5): In this variant the weights w_{ij} are equal to 1 and θ, γ and β are different. To enhance the exploration of the network to sampled nodes which are further away from the previous visited nodes, we consider the case $\theta < \gamma < \beta$. The stationary distributions in this case are not known and we will use the empirical distribution obtained through simulations.

Node2vec-6 (N2V-6): This is the more general variant extending N2V-5 to have weights. Again, the most interesting case is $\theta < \gamma < \beta$. As in N2V-5 the stationary distributions are unknown. However, we include this sampling scheme for a full evaluation of the performance of N2V. We believe that for the network model an approximation can be obtained for stationary distributions through the resolvability of the model classes $\mathscr{P}$ and $\mathscr{U}$. Due to the technical nature of the problem, it is outside the scope of this paper, and will be considered in a future work.

For comparison to RWs, we will also use the following baseline samplings. These can be viewed as "ideal" for sampling purposes and correspond to the limiting distributions of some RWs.

Node Sampling (NS). NS sampling requires full access to the network and is unavailable for many real networks. In the classical NS, nodes are chosen independently and uniformly from the network with replacement.

Edge Sampling (ES). In the classical ES, edges are chosen independently and uniformly from the network with replacement. Since ES selects edges rather than nodes to populate the sample, the node set is constructed by including both incident nodes in the sample when a particular edge is sampled.

Estimation of Attribute Distributions and Homophily Measures

We consider here estimation in the case of discrete-valued attributes; the case of continuous-valued attributes is discussed at the end of this work. Our estimators of quantities of interest will be based on one of the following two general estimators. The first estimator is for the proportion $p(A)$ of nodes i with a certain characteristic $A(i)$ taking value A. The characteristic takes discrete values and could be the discrete attribute $a_i = a(i)$ itself, the degree $d_i = d(i)$, the combination of the latter two, etc. The estimator of $p(A)$ for a random walk is defined as follows. Run a random walk (any of the sampling schemes described above) for n steps and let i_s denote the sth node sampled by the random walk, for $1 \leq s \leq n$. Since nodes are sampled with replacement and with probabilities π_i in the stationary regime, the proportion $p(A)$ can be estimated as

$$\widehat{p}(A) = \frac{1}{Nn} \sum_{s=1}^{n} \frac{\mathbf{1}\{A(i_s) = A\}}{\pi_{i_s}}, \tag{12}$$

where $\mathbf{1}\{E\} = 1$ if E is true and 0 otherwise [27] (Chapter 5). If the total number of nodes N is unknown, its estimator is given by $\widehat{N} = (1/n)\sum_s 1/\pi_{i_s}$, and (12) becomes

$$\widehat{p}(A) = \frac{1}{\sum_{s=1}^{n} 1/\pi_{i_s}} \sum_{s=1}^{n} \frac{\mathbf{1}\{A(i_s) = A\}}{\pi_{i_s}}. \tag{13}$$

A direct application of e.g. (12) yields the following estimators for the proportion $p(k, a)$ of nodes with degree k and attribute a, the proportion $p(a)$ of nodes with attribute a, and the conditional proportion $p(k|a) = p(k, a)/p(a)$ of nodes of degree k having attribute a:

$$\widehat{p}(k, a) = \frac{1}{Nn} \sum_{s=1}^{n} \frac{\mathbf{1}\{d(i_s) = k, a(i_s) = a\}}{\pi_{i_s}}, \qquad a \in \mathcal{A}, \tag{14}$$

$$\widehat{p}(a) = \frac{1}{Nn} \sum_{s=1}^{n} \frac{\mathbf{1}\{a(i_s) = a\}}{\pi_{i_s}}, \qquad a \in \mathcal{A}, \tag{15}$$

$$\widehat{p}(k|a) = \sum_{s=1}^{n} \frac{\mathbf{1}\{d(i_s) = k, a(i_s) = a\}}{\pi_{i_s}} \Big/ \sum_{s=1}^{n} \frac{\mathbf{1}\{a(i_s) = a\}}{\pi_{i_s}}, \qquad a \in \mathcal{A}. \tag{16}$$

We note that the quantities in (14)–(16) are given in terms of the sample obtained through the random walk used with N estimated by $\widehat{N}$.

The performance of $\widehat{p}(A)$ in (12) and hence the components of the estimators (14)–(16) can be assessed through their MSE. For fixed A, the MSE of $\widehat{p}(A)$ is given by $E[(\widehat{p}(A) - p(A))^2]$. In the stationary regime, $\widehat{p}(A)$ in (12) is an unbiased estimator of $p(A)$ and the MSE is equal to the variance $V[\widehat{p}(A)]$. The variance of $\widehat{p}(A)$ can be related to the spectral gap of the RW. More specifically, let P be the associated transition matrix of the random walk with eigenvalues (real by reversibility): $1 = \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_N \geq -1$. The spectral gap is defined as

$\delta = 1 - \lambda_2$. Equivalently, the relaxation time of the RW is the reciprocal of the spectral gap. A larger spectral gap implies a faster convergence of the RW to its stationary distribution. From [28] (Proposition 4.29), we have

$$V(\widehat{p}(A)) \leq \frac{2\Lambda(A)}{\delta n}\left(1 + \frac{\delta}{2n}\right), \tag{17}$$

where $\Lambda(A) = \sum_{i=1}^{N} \mathbf{1}\{A(i) = A\}/(N^2\pi_i)$. The error in estimating the proportion of nodes with characteristic A is thus proportional to the inverse of the spectral gap and $\Lambda(A)$; the latter is small if the probability of sampling nodes with characteristic A is large. We will see in Section Experiments that for N2V-2, if edge weights w_{ij} are *inversely* related to the concordance of the attributes, thus encouraging the walk to explore vertices with different attributes, then in some settings, this increases δ and decreases $\Lambda(a)$ (for attributes with small proportions), resulting in a smaller variance of the estimator for the proportion $p(a)$ of nodes with attribute a.

The second estimator is for the proportion $p(B)$ of edges (i, j) with a certain characteristic $B(i, j)$ taking value B. The values B are assumed to be discrete. For the random walk considered above, since edges are sampled with probabilities π_{ij} in the stationary regime, the proportion $p(B)$ can be estimated similarly to (12) as

$$\widehat{p}(B) = \frac{1}{(n-1)|\mathcal{E}|}\sum_{s=1}^{n-1}\frac{\mathbf{1}\{B(i_s, i_{s+1}) = B\}}{\pi_{i_s,i_{s+1}}} \tag{18}$$

and if needed, the number of edges as

$$\widehat{|\mathcal{E}|} = \frac{1}{n-1}\sum_{s=1}^{n-1}\frac{1}{\pi_{i_s,i_{s+1}}}. \tag{19}$$

A direct application of (18)–(19) is to estimation of homophily measures D_a and $H_{aa'}$ in (2) and (3) as:

$$\widehat{D}_a = \widehat{|\mathcal{E}_{aa}|}\Big/\left(\binom{\widehat{|\mathcal{V}_a|}}{2}\widehat{p}\right), \qquad \widehat{H}_{aa'} = \widehat{|\mathcal{E}_{aa'}|}\Big/\left(\widehat{|\mathcal{V}_a|}\widehat{|\mathcal{V}_{a'}|}\widehat{p}\right), \tag{20}$$

where $\widehat{|\mathcal{V}_a|} = \widehat{N}\widehat{p}(a)$, $\widehat{p} = \widehat{|\mathcal{E}|}/\binom{\widehat{|N|}}{2}$ and

$$\widehat{|\mathcal{E}_{aa'}|} = \frac{1}{n-1}\sum_{s=1}^{n-1}\frac{\mathbf{1}\{(a(i_s), a(i_{s+1})) = (a, a') \vee (a(i_s), a(i_{s+1})) = (a', a)\}}{\pi_{i_s,i_{s+1}}}, \tag{21}$$

where $a, a' \in \mathcal{A}$. We note again that the quantities in (19)–(21) are given by the sample obtained through the respective random walk used. We are not aware of the results of the type (17) to assess the variability of the estimator $\widehat{p}(B)$ in (18).

In terms of complexity of the learning framework, the random walks considered in this work are computationally efficient in terms of both space and time requirements [18]. For instance, for each visited node, we need to check the immediate neighbors and their attributes. For the second order random walks (N2V-3, -4, -5 and -6), we

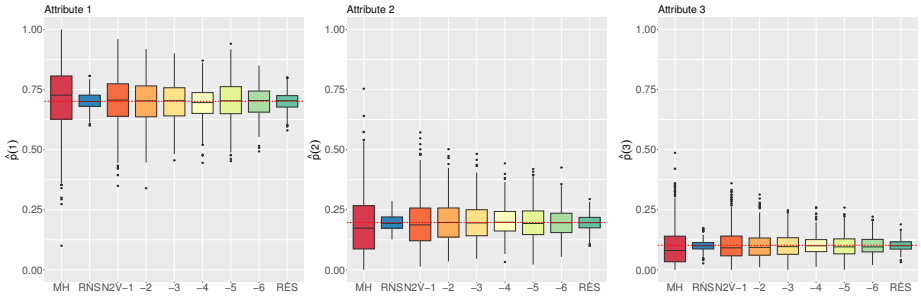


Figure 5 Boxplots of 500 estimates of the attribute distribution under model class $\mathscr{P}(\alpha, \mu, f)$ with homophily where $N = 2000$, $\alpha = 0.2$, $\mu = (0.7, 0.2, 0.1)$, $f(a, a) = 0.8$, $f(a, a') = 0.1$, $a \neq a' = 1, 2, 3$ and $m = 2$. The sample size for each sampling method is $0.15N$. N2V-2 ($w_{ij} = 0.3$ if nodes have equal attributes and 1 otherwise); N2V-3 ($\theta = 10^{-3}$, $\beta = \gamma = 1$), N2V-4 ($\theta = 10^{-3}$, $\beta = \gamma = 1$, $w_{ij} = 0.3$ if nodes have equal attributes and 1 otherwise), N2V-5 ($\theta = 10^{-3}$, $\gamma = 0.1, \beta = 1$), N2V-6 ($\theta w_{ij} = 10^{-3}$, $\gamma w_{ij} = 0.1$, and $\beta w_{ij} = 0.3$ if nodes have equal attributes and 1 otherwise). The red dotted lines represent the true values $p(a)$, $a = 1, 2, 3$.

need additionally to keep track of the interconnections between the neighbors of the current visited node, however, the average degree of the graph is usually small for most real world networks. The proposed estimators are obtained from simple weighted sample statistics.

Experiments

In this section, we assess the performance of the sampling methods and estimators in learning the attribute distribution, degree distribution per attribute and homophily measures on synthetic and real-world networks with discrete attributes.

Synthetic Network with Homophily

We consider the model class $\mathscr{P}(\alpha, \mu, f)$ with $N = 2000$ nodes and 3 discrete attributes. In the generation of the network, each node that enters the system has attribute 1, 2 or 3 with probabilities $\mu_1 = 0.7$, $\mu_2 = 0.2$, $\mu_3 = 0.1$, respectively, and connects to $m = 2$ nodes proportional to (1), where $f(a, a) = 0.8$, $f(a, a') = 0.1$, $a, a' = 1, 2, 3$, $a \neq a'$. We investigate the effect of homophily in the estimation of the quantities of interest in a controlled environment for the two most interesting network topologies: sublinear ($\alpha = 0.2$) and linear ($\alpha = 1$).

Attribute Distribution

Setting 1 ($\alpha = 0.2$): The evaluation of the several sampling methods in learning the attribute distribution using (15) assuming N unknown is shown in Fig. 5. Each boxplot is constructed from the results of 500 estimates. The length of each walk is $0.15N$. MHRW has an important property that the stationary distribution is uniform over all the nodes. Thus, in principle, MHRW is equivalent to RNS of the network for an infinite RW. In practice, MHRW typically requires sample sizes of $O(N)$ to achieve the stationary distribution [29]. It is challenging to use MHRW for large scale networks with millions of nodes, where typical sample size is much smaller than the network size. Networks with a strong homophily are problematic in this case since MHRW tends to get stuck in nodes with the same attributes. The

Table 4 The variation of spectral gap (δ) and $\Lambda(3)$ from the bound of the variance of $\widehat{p}(3)$ under model class $\mathscr{P}(\alpha, \mu, f)$ and sampling method parameters as described in Fig. 5.

	MH	N2V-1	N2V-2	N2V-3	N2V-4	N2V-5	N2V-6
spectral gap (δ)	0.064	0.137	0.110	0.359	0.421	0.380	0.420
$\Lambda(3)$	0.102	0.149	0.116	0.149	0.120	0.150	0.120

Table 5 Empirical standard deviation of $\widehat{p}(3)$, and the variation of spectral gap (δ) and $\Lambda(3)$ (from the bound of the variance of $\widehat{p}(3)$) with N2V-2 ($w_{ij} = 1, a(i) \neq a(j)$) under model class $\mathscr{P}(\alpha, \mu, f)$ where $N = 2000$, $\alpha = 0.2$, $\mu = (0.7, 0.2, 0.1)$, $f(a, a) = 0.8$, $f(a, a') = 0.1$, $a \neq a' = 1, 2, 3$ and $m = 2$.

$w_{ij}, a(i) = a(j)$	0.05	0.25	0.5	0.75	1	1.25	1.5
s.d.	0.070	0.054	0.058	0.059	0.062	0.073	0.075
spectral gap (δ)	0.027	0.098	0.1344	0.139	0.137	0.130	0.122
$\Lambda(3)$	0.142	0.115	0.124	0.137	0.149	0.160	0.171

classical variant of node2vec, N2V-1, which like MHRW is also attribute agnostic has the property that the stationary distribution is uniform over all the edges. N2V-1 is equivalent to RES of the network for an infinite RW. In practice, it suffers from the same drawbacks of MHRW to a lower extent. The poor performance can also be explained through the bound of the variance (17). Table 4 shows that MHRW has the lowest spectral gap while N2V-1 has a high value $\Lambda(3)$ for attribute 3 (this is detailed next for N2V-2).

The attribute aware samplers like N2V-2 use node attribute to determine the next node to add to the sample, by checking the attribute of the node against the attribute of the last node added to the sample. To simplify the exposition (instead of w_{ij} for nodes i and j), we write $\overline{w}_{aa}$ for the weights of nodes with the same attributes, and $\overline{w}_{aa'}$ with different attributes. Table 5 shows the effects of the weights in the standard deviation of the estimate for N2V-2 for attribute 3. To explain their differences, we turn to the bound of the variance of the estimator (17). The error in estimating the proportion of nodes with an attribute a is upper bounded by the inverse of the spectral gap. If $\overline{w}_{aa}$ is much smaller than $\overline{w}_{aa'} = 1$, say $\overline{w}_{aa} = 0.05$, then the movements of N2V-2 between different node attributes are very frequent and exploration within each attribute is insufficient. In this case, the spectral gap is low creating a bottleneck for approaching the stationary probability. As $\overline{w}_{aa}$ increases the inter-attribute moves are less frequent, accelerating the convergence to the stationary distribution. On the other hand, when $\overline{w}_{aa}$ becomes greater or equal than $\overline{w}_{aa'}$, the spectral gap decreases until that N2V-2 hardly transits from one attribute value to another. The error in estimating the attribute distribution is also bounded by the quantity $\Lambda(a)$ which is small if the probability of sampling nodes with attribute a is large. We also observe from Table 5 the effect of $\overline{w}_{aa}$ on the value $\Lambda(a)$ for attribute 3. The tradeoff between δ and $\Lambda(a)$ explains the smaller standard deviation for attribute 3 of N2V-2 with $\overline{w}_{aa} = 0.25$. The convex behavior of the empirical standard deviation as a function of $\overline{w}_{aa}$ will be explored at the end of this work in the guidelines for setting the weights of attribute aware samplers.

In N2V-3, the parameter θ of the propensity for the random walk to backtrack is set close to zero $\theta = 10^{-3}$ such that if the walker arrives at a node with degree 1, it always backtracks in the next time step since this is the only possible move, and $\beta = \gamma = 1$. In this case, N2V-3 tends to explore better the network, avoiding the redundancy of nodes in the sample which accelerates the convergence (see the spectral gap in Table 4). The result is consistent with the non-backtracking RWs on

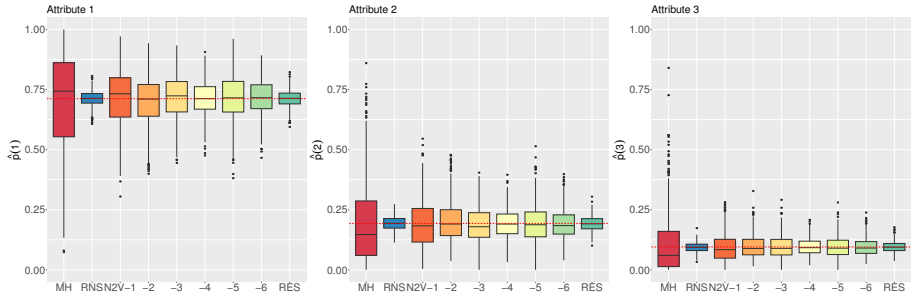


Figure 6 Boxplots of 500 estimates of the attribute distribution under model class $\mathscr{P}(\alpha, \mu, f)$ with $\alpha = 1$. See Fig. 5 for the remaining parameters of the model and sampling methods.

regular graphs [30]. In many cases, they find spectral gap "twice as good" compared to the classical RW, as also in our case.

N2V-4 combines features of both attribute aware and non-backtracking samplers. We use the same weights and backtracking parameters as in N2V-2 and N2V-3 above. Since the stationary distribution π_i in (15) is not known, it is obtained through simulations. The results show that N2V-4 can provide better estimates with lower variability compared to N2V-2 and N2V-3. This can be explained by the increase of the spectral gap while keeping $\Lambda(a)$ small for attribute values 2 and 3 (see Table 4). We have confirmed the use of the approximation in (11) for the stationary distribution of N2V-4. The choice is heuristic but the results show very good accuracy compared to the empirical distribution for this network scenario.

N2V-5 ignores the attributes of nodes while sampling the network. We set $\theta = 10^{-3}$, $\gamma = 0.1$ and $\beta = 1$, forcing the RW to explore non-common neighbors of the previous and currently visited nodes. The performance is worse compared with N2V-4 with the decrease of the spectral gap and the increase of $\Lambda(3)$ (Table 4). N2V-6 is the version of N2V-5 with attribute aware sampling. We now set $\beta w_{ij} = 0.3$ if nodes have equal attributes and 1 otherwise as in N2V-4 and keep the other parameters used in N2V-5. There is an improvement of performance, however, its variability is similar to N2V-4. In both N2V-5 and -6, the stationary distributions used in the estimation are obtained through simulations.

Setting 2 ($\alpha = 1$): We next consider the linear model class $\mathscr{P}(1, \mu, f)$ case, where μ, f, N and the sampling rate are the same as in Setting 1. The boxplots of 500 estimates for each sampling scheme using (15) are given in Fig. 6. In this case, the performance of MHRW is worse due to the existent of high degree nodes which tend to be avoided by MHRW, reducing the spectral gap. Note that high degree vertices increase "conductance" in the network (small world phenomenon) and hence avoiding them decreases the mixing time of MHRW. For the variants of N2V the estimates for attributes 2 and 3 tend to be better. This can be explained by the homophily and preferential attachment in the model which enables different types of attachment propensities as we now indicate. The attributes with small proportions 2 and 3 will be mainly attracted by the same node attributes. However, due to the preferential attachment, nodes with attributes from small proportions will also be partly attracted to the majority proportion of nodes with attribute 1 (see Fig. 1-b).

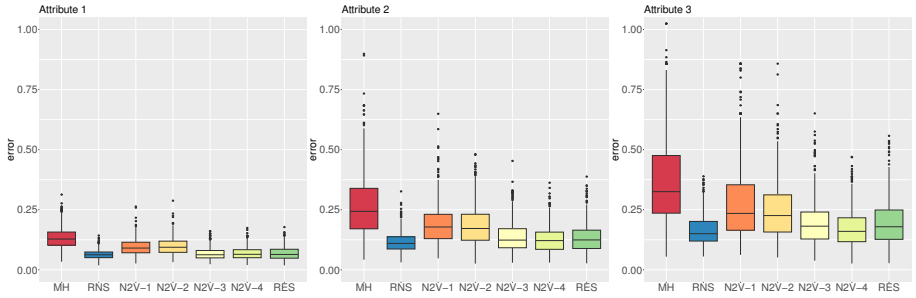


Figure 7 Boxplots of 500 estimation errors of the degree distribution per attribute under model class $\mathscr{P}(\alpha, \mu, f)$ with homophily where $N = 2000$, $\alpha = 0.2$, $\mu = (0.7, 0.2, 0.1)$, $f(a, a) = 0.8$, $f(a, a') = 0.1$, $a \neq a' = 1, 2, 3$ and $m = 2$. The sample size for each sampling method is $0.2N$. N2V-2 ($w_{ij} = 0.3$ if nodes have equal attributes and 1 otherwise); N2V-3 ($\theta = 10^{-3}$, $\beta = \gamma = 1$), N2V-4 ($\theta = 10^{-3}$, $\beta = \gamma = 1$, $w_{ij} = 0.3$ if nodes have equal attributes and 1 otherwise).

Therefore, the variability in the estimation tends to be smaller for attributes with lower proportions. The ranking of the performance of sampling methods is the same as in the sublinear case.

Other settings such as the presence of weak homophily and balanced attributes, i.e. the distribution of attributes in the network being uniform will be investigated with real data.

Degree Distribution per Attribute

Setting 3 ($\alpha = 0.2$): Fig. 7 depicts the boxplots of the estimation error $(\sum_k (\widehat{p}(k|a) - p(k|a))^2)^{1/2}$ of the degree distribution per attribute for a sublinear network from 500 estimates under MHRW, N2V-1 to -4, and baseline sampling methods. Since the stationary distributions of N2V-5 and -6 are not known and the N2V-5 and -6 performances approach N2V-3 and -4, respectively, we omitted them in the plot. The number of nodes sampled is $0.2N$ and the parameters of N2V-3 and 4 are the same as in Setting 1. N2V-4 achieves the highest performance especially for attributes 2 and 3 (even compared with RES) due to being attribute aware. We use its empirical stationary distribution and also check the approximation (11) which shows similar boxplots. On the other hand, MHRW has a poor performance compared with the baseline RNS. The results for the variants of N2V are consistent with the estimation of the attribute distribution.

Homophily Measures

Setting 4 ($\alpha = 1$): The homophily measures are $D_1 = 1.34$, $D_2 = 3.44$, $D_3 = 4.87$, $H_{12} = 0.28$, $H_{13} = 0.37$, $H_{23} = 0.56$. Fig. 8 shows the estimates of the dyadicity and heterophilicity using N2V variants with known or approximate stationary distribution. The estimators in (20) involves the ratio of several quantities which are sensitive to small deviations. Thus a larger sample size $0.3N$ is used to reduce the variability. The other parameters are the same as in Setting 2. We have omitted MHRW in the plots due to having the worst performance and also the baseline RNS. For the heterophilicity measure, N2V-4 achieves the lower variability followed by RES. We note that $\widehat{H}_{aa'}$ in (20) involves the estimation of the number of edges

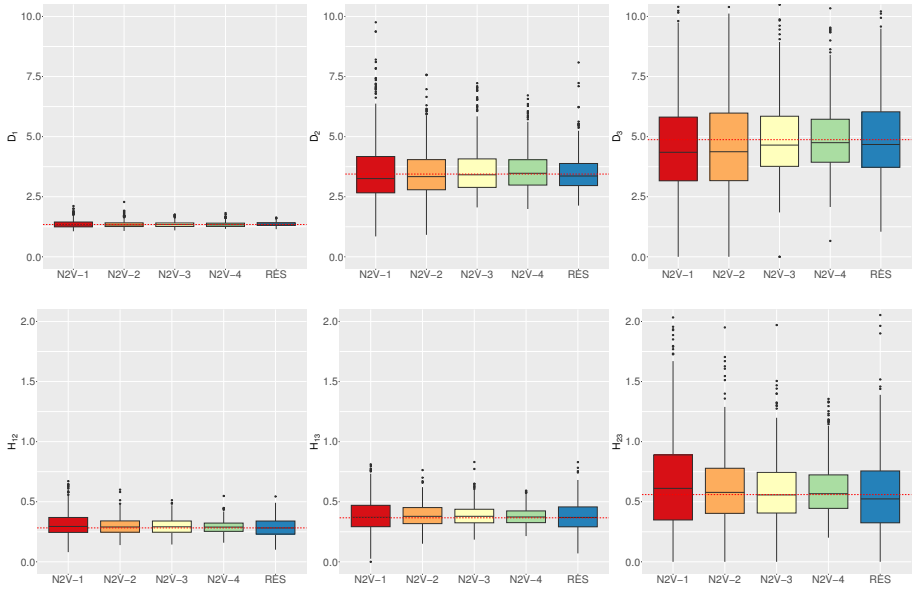


Figure 8 Boxplots of 500 estimates of homophily measures under model class $\mathscr{P}(\alpha, \mu, f)$ with homophily where $N = 2000$, $\alpha = 1$, $\mu = (0.7, 0.2, 0.1)$, $f(a, a) = 0.8$, $f(a, a') = 0.1$, $a \neq a' = 1, 2, 3$ and $m = 2$. The sample size for each sampling method is $0.3N$. N2V-2 ($w_{ij} = 0.3$ if nodes have equal attributes and 1 otherwise), N2V-3 ($\theta = 10^{-3}$, $\beta = \gamma = 1$), N2V-4 ($\theta = 10^{-3}$, $\beta = \gamma = 1$, $w_{ij} = 0.3$ if nodes have equal attributes and 1 otherwise). The red dotted lines represent the true values.

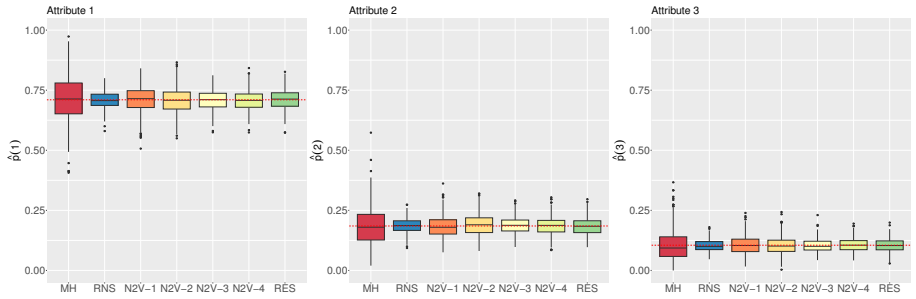


Figure 9 Boxplots of the estimates of the attribute distribution under model class $\mathscr{P}(\alpha, \mu, f)$ with heterophily where $N = 2000$, $\alpha = 1$, $\mu = (0.7, 0.2, 0.1)$, $f(a, a) = 0.2$, $f(a, a') = 0.4$, $a \neq a' = 1, 2, 3$ and $m = 2$. The sample size for each sampling method is $0.15N$. N2V-2 ($w_{ij} = 1$ if nodes have equal attributes and 0.8 otherwise), N2V-3 ($\theta = 10^{-3}$, $\beta = \gamma = 1$), N2V-4 ($\theta = 10^{-3}$, $\beta = \gamma = 1$, $w_{ij} = 1$ if nodes have different attributes and 0.8 otherwise). The red dotted lines represent the true values.

between different attribute nodes, which due to the reduced number of these connections is better estimated with N2V-4 than RES.

Synthetic Network with Heterophily

Attribute Distribution

Setting 6: We consider the model class $\mathscr{P}(\alpha = 1, \mu = (0.7, 0.3, 0.1), f)$ with $f(a, a) = 0.2$, $f(a, a') = 0.4$, $a, a' = 1, 2, 3$, $a \neq a'$. The network size and sampling rate are the same as in the synthetic network with homophily (Settings 1 and

Table 6 Empirical networks characteristics (total number of nodes and edges, attribute types, dyadicity and heterophilicity measures).

	N	$\|\mathcal{E}\|$	attribute	D_1	D_2	H_{12}
Wikipedia	1595	2809	male/female	1.0810	1.706	0.710
Blogs	1222	16714	right/left	1.733	1.901	0.189
APS	1281	3064	subfield 1/2	1.491	2.487	0.128
Swarthmore	1517	53725	male/female	1.082	1.052	0.933

2). The network generated is heterophilic with measures $D_1 = 0.667$, $D_2 = 0.573$, $D_3 = 0.962$, $H_{12} = 1.261$, $H_{13} = 1.668$ and $H_{23} = 1.378$. Fig. 9 gives the estimates of the attribute distribution under several sampling schemes using (15). With heterophily, for attributes aware samplers the weights are higher if nodes have equal attributes. For N2V-2 and N2V-4 the weights are $w_{ij} = 1$ if nodes have different attributes and 0.8 otherwise. The differences between the different sampling methods are now smaller. In this case, even though most edges are heterophilic, networks will also contain edges between nodes of the same attribute type (see Fig. 2-b). This is specially true for nodes with attribute 1 where locally they connect to few other nodes with attribute 1, but globally there are many connections between them. This mixing of different types of edges explains why heterophilic networks can achieve high overall performance among the different sampling methods. The spectral gap of the random walks increases and also the quantity $\Lambda(.)$ decreases which also explains the results.

Empirical Networks

We analyze four publicly available datasets of real attributed networks from different domains and with different homophily levels. Table 6 shows some key characteristics of interest. Wikipedia dataset is a hyperlink network where nodes represent U.S. politicians with attributes as either male or female. Blogs dataset is a network from political blogs from the 2004 U.S. election. Nodes represent blog pages and edges hyper-links between them. Each blog is either right- or left-leaning as attribute. APS is a scientific network from the American Physical Society where nodes represent articles from two subfields and edges represent citations. Swarthmore is a university network with friendship links between users' pages with attribute gender (male or female). The estimation of the quantities of interest below are replicated 500 times for each sampling scheme.

Attribute Distribution

Fig. 10 shows the results of estimation of the attribute distributions using (15) for all data sets. We investigate only the sampling methods with known (or approximately computable) stationary distributions. For N2V-3 and -4, we use $\theta = 10^{-3}$, $\gamma = \beta = 1$ through this section. The sample size is $0.15N$. Wikipedia has unbalanced attributes and moderate homophily. For N2V-2 and -4 the weights are $w_{ij} = 0.75$ if nodes have equal attributes and 1 otherwise. The performance of MHRW with real data shows again the worst performance. The variants 3 and 4 of N2V presents the lowest variability. Blogs is an approximately balanced attribute data set with a significant homophily. The weights for the variants of N2V are $w_{ij} = 0.3$ if nodes have equal attributes and 1 otherwise. Due to the high density of edges (i.e., the fraction of existing edges out of all possible edges, $|\mathcal{E}|/\binom{N}{2}$) the performance of

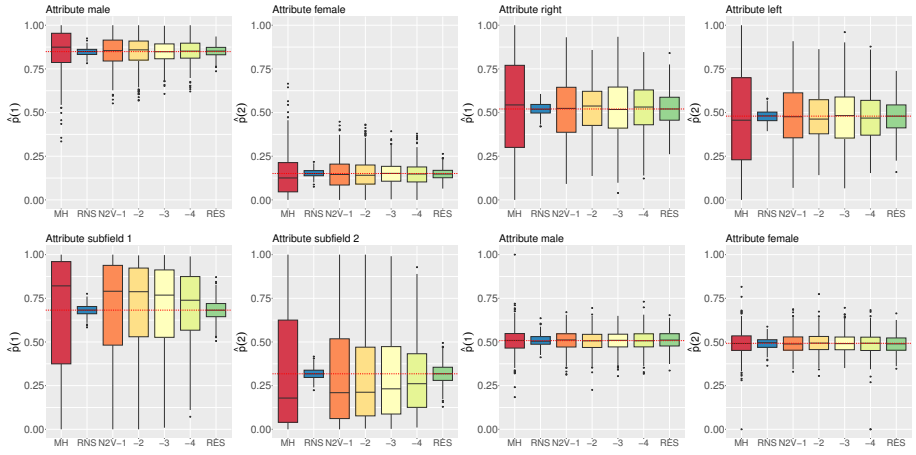


Figure 10 Boxplots of 500 estimates of the attribute distribution: Wikipedia (N2V-2 and -4: $w_{ij} = 0.75$ if nodes have equal attributes and 1 otherwise), Blogs (N2V-2 and -4: $w_{ij} = 0.3$ if nodes have equal attributes and 1 otherwise), APS (N2V-2 and -4: $w_{ij} = 0.25$ if nodes have equal attributes and 1 otherwise), Swarthmore (N2V-2 and -4: $w_{ij} = 0.95$ if nodes have equal attributes and 1 otherwise). N2V-3 and -4, $\theta = 10^{-3}$, $\gamma = \beta = 1$. The sample size for each sampling method is $0.15N$. The red dotted lines represent the true values.

N2V-3 is similar to N2V-1. APS is an unbalanced attribute dataset with strong homophily. In this case $w_{ij} = 0.25$ if nodes have equal attributes and 1 otherwise. Swarthmore is a dataset which is very weakly homophilic. The diferences between the sampling methods are less significant where we use $w_{ij} = 0.95$ if nodes have equal attributes and 1 otherwise. These empirical networks are heterogenous with respect to homophily complementing the settings considered in the synthetic case. In [19] we have estimated the attribute distribution of a Facebook webgraph dataset restricted to pages from four attributes (politicians, governmental organizations, television shows and companies) where edges represent mutual likes between sites.

Degree Distribution per Attribute

Fig. 11 depicts the estimation error $(\sum_k (\widehat{p}(k|a) - p(k|a))^2)^{1/2}$ of the degree distribution for each attribute of Wikipedia and APS. The parameters for the different sampling methods are the same as for the estimation of attribute distribution with sample size $0.2N$. The degree distributions for both attributes are heavy-tailed in the two datasets. For the majority attributes the tail exponents are 2.823 and 3, respectively, for Wikipedia and Blogs. The error in the estimation decreases significantly with N2V-4, especially for the minority attribute.

Homophily Measures

The dyadicity and heterophilicity measures using (20) are given in Fig. 12 for Wikipedia and Blogs. Only N2V variants have been considered in the evaluation with the same parameters as above and sample size $0.3N$. The performance of the samplings methods for Wikipedia are in line with the synthetic model with discrete attribute set. The high density of edges in Blogs, as discussed above, explains the inferior performance of N2V-3 similar to N2V-1 especially in the estimation of H_{12}.

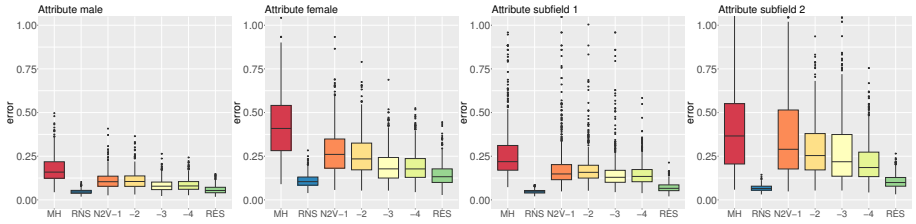


Figure 11 Boxplots of the estimation error of the degree distribution per attribute: Wikipedia (N2V-2 and -4: $w_{ij} = 0.75$ if nodes have equal attributes and 1 otherwise), APS (N2V-2 and -4: $w_{ij} = 0.25$ if nodes have equal attributes and 1 otherwise). N2V-3 and -4, $\theta = 10^{-3}$, $\gamma = \beta = 1$. The sample size for each sampling method is $0.2N$.

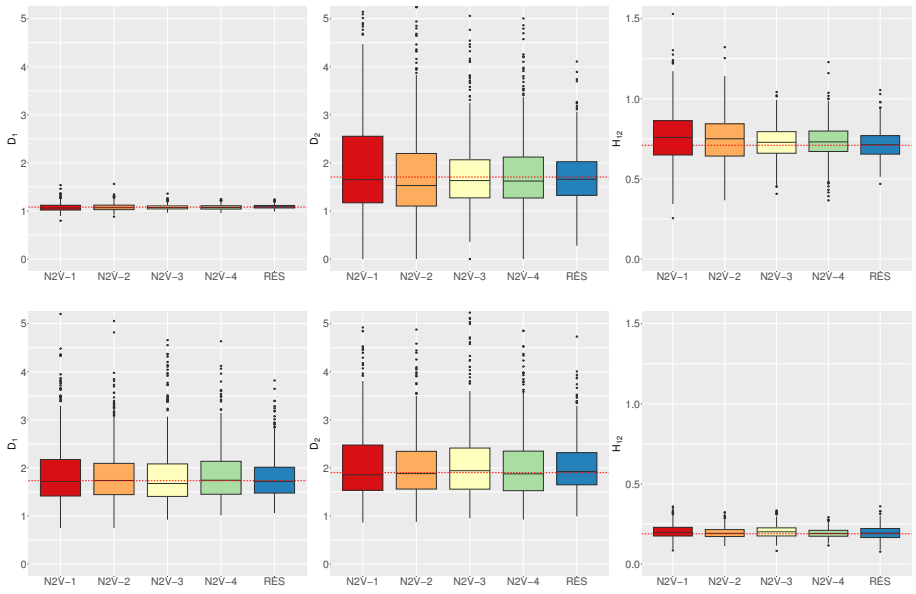


Figure 12 Boxplots of the estimates of homophily measures: Wikipedia (N2V-2 and -4: $w_{ij} = 0.75$ if nodes have equal attributes and 1 otherwise), Blogs (N2V-2 and -4: $w_{ij} = 0.3$ if nodes have equal attributes and 1 otherwise). N2V-3 and -4, $\theta = 10^{-3}$, $\gamma = \beta = 1$. The sample size for each sampling method is $0.3N$. The red dotted lines represent the true values.

Extensions and Future Directions

How to Sample the Network and set the Sampling Method Parameters?

Here are some guidelines on how to sample and learn the attribute functionals of a network. If the homophily level is unknown (or even if it is not known if the network is homophilic), the network should be sampled with N2V-3 to estimate the dyadicity and heterophilicity measures. As seen from our experiments the backtracking parameter should be close to zero and the other parameters equal to one. In the case that the sampled network indicates that the network is homophilic, we propose the following approach to set the initial edge weights of attribute aware samplers (N2V-2 and N2V-4) to estimate the attribute distribution (and additionally the degree distribution). If dyadicity is, say, greater than 1.5 and heterophilicity is less than 0.5, then set the weights to $w_{ij} = 0.3$ if nodes have equal attributes and 1 otherwise. For lower homophily levels, set $w_{ij} = 0.7$ if nodes have equal attributes.

(In the case of N2V-4, additionally the backtracking parameter should be close to zero.) As observed in the Section Synthetic Network with Homophily (Setting 1), the empirical standard deviation of the estimator of the attribute distribution as a function of the weights w_{ij} (when nodes have equal attributes) is convex. Thus, the weights can then be tuned in practice as follows if feasible. (1) Fix the initial set of weights as described above and a minority attribute, and run n (say, greater than 10) independent attribute aware samplers for a number of steps and obtain the empirical standard deviation of the n estimates of the proportion of the minority attribute; (2) The weights of the n samplers are then increased (decreased) with increment Δ and run again to compute the empirical standard deviation; (3) The previous step is repeated until an inflection point of the empirical standard deviation is reached and the "optimal" weight is outputted.

Continuous Attributes

The estimators (12) and (18) were defined for node and edge characteristics that are discrete. But they have natural continuous analogues. More specifically, in connection to (12), assume that the characteristic $A(i)$ values are such that $A \in \mathbb{R}^d$. Then, we expect the density $g(A)$ to be estimated by the kernel smoothing as

$$\widehat{g}(A) = \sum_{s=1}^{n} K\left(\frac{A - A(i_s)}{h}\right) \frac{1}{h^d} \widetilde{w}_s, \tag{22}$$

where $h > 0$ is a bandwidth, $K : \mathbb{R}^d \to \mathbb{R}$ is a kernel function, and the weights $\widetilde{w}_s$ satisfy

$$\widetilde{w}_s \propto \frac{1}{\pi_{i_s}}, \quad \sum_{s=1}^{n} \widetilde{w}_s = 1. \tag{23}$$

For the density $g(a)$ of continuous attributes $a(i) \in \mathbb{R}$, the estimator (22) was explored briefly in synthetic and real networks in our conference paper [19].

Similarly, the continuous analogue of (18) is

$$\widehat{g}(B) = \sum_{s=1}^{n-1} K\left(\frac{B - B(i_s, i_{s+1})}{h}\right) \frac{1}{h^d} \widetilde{w}_{s,s+1}, \tag{24}$$

where K and h are as in (22), and the weights $\widetilde{w}_{s,s+1}$ satisfy

$$\widetilde{w}_{s,s+1} \propto \frac{1}{\pi_{i_s,i_{s+1}}}, \quad \sum_{s=1}^{n-1} \widetilde{w}_{s,s+1} = 1. \tag{25}$$

Exploring (22) and (24) further is left for future work. For the attribute aware samplers the weights can be taken as $w_{ij} = |a(i) - a(j)|^b$, which allows moving between similar attribute values of nodes but also giving more weight to edges with different values. The choice of b is motivated by similar arguments as in the case of discrete attributes. If the weights between edges of different groups are too large, then the convergence is decelerated because exploration within the same group attribute is not sufficient due to the inter-group moves.

Future Directions

We expect to show that for the parameter $m \geq 2$ in model class $\mathscr{P}$, the networks are 'expanders' in the sense that the mixing time of RWs on the network is of a much smaller order than the network size (typically logarithmic in network size) [31, 32]. This would indicate that, although explicitly finding the stationary distribution is infeasible in most cases (e.g. in N2V-4,5,6 discussed above), it can be approximated by observing the RW for a relatively small number of steps. A description of the local limits of neighborhoods of typical vertices in the network [33, 34, 35] will then provide tractable recursive distributional equations (e.g. [36] for Pagerank distribution) characterizing the limiting empirical stationary distribution of the RW (as the network size grows). This representation can be exploited to analyze detailed behavior of this limiting distribution including tail exponents, means, etc.

Random walks are also closely tied to ranking mechanisms such as the Pagerank centrality, and we plan to study the impact of the parameters driving the random walk on such centrality scores, thus looping back to one of the central motivations for studying attributed networks namely fairness of ranking mechanisms [12]. Other questions, including learning joint distributions of the multivariate attribute distributions, both in terms of developing synthetic models, as well as real world data will also be considered. We considered simple time snapshots of the network process, without directionality information, for estimation in this work, but in future work it will be interesting to exploit the temporality and directionality in network data. Finally, there has been significant recent interest in incorporating higher order interactions (network data and models largely hinge on binary or pairwise interactions) in the evolution of networks and the impact of dynamics such as percolation and epidemics resulting from such interactions [37, 38, 39, 40, 41, 42]. Exploring versions of such questions incorporating attribute information suggests fascinating new directions of research.

Conclusions

In this paper, we developed a statistical framework for learning attribute functionals through sampling in networks with homophily. First, we proposed a generalization of the preferential attachment model with homophily (model class $\mathscr{P}$). We described a related model (model class $\mathscr{U}$), that is significantly more amenable to analysis, formalizing the notion of *resolvability*, which provides explicit information (degree distribution of an attribute, homophily and heterophily statistics) for model class $\mathscr{P}$ by using model class $\mathscr{U}$. Second, we introduced link trace samplers (random walks) with weights for networks with restricted access that explore better the attribute space (attributed aware). Third, estimators that correct the bias of the considered sampler methods were proposed for the several attribute and geometric quantities of interest. Fourth, we showed experimental results for synthetic (using model class $\mathscr{P}$) and a variety of real world datasets, demonstrating that attribute aware samplers are more efficient and outperform attribute agnostic random walks samplers for several network settings. Finally, we presented extensions of the developed framework including continuous attributes and directions for future work.

Acknowledgements
The authors would like to thank the editors and two anonymous reviewers for their comments that led to significant improvements in the paper.

Funding
S. Banerjee is partially supported by the NSF CAREER award DMS-2141621. S. Bhamidi and V. Pipiras are partially supported by NSF DMS-2113662. S. Banerjee, S. Bhamidi and V.Pipiras are partially supported by NSF RTG grant DMS-2134107.

Availability of data and materials
All data are publicly available on GitHub at GESIS – Leibniz Institute for the Social Sciences: https://github.com/orgs/gesiscss/repositories.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
All authors contributed equally to the manuscript.

Author details
[1]Center for Computational and Stochastic Mathematics, University of Lisbon, Avenida Rovisco Pais, 1049-001, Lisbon, Portugal. [2]Department of Statistics and Operations Research, University of North Carolina, CB 3260, Chapel Hill, NC 27599, USA.

References

1. Fan, H., Zhong, Y., Zeng, G., Sun, L.: Attributed network representation learning via improved graph attention with robust negative sampling. Applied Intelligence **51**(1), 416–426 (2021)
2. Chang, C.-H., Chang, C.-S., Chang, C.-T., Lee, D.-S., Lu, P.-E.: Exponentially twisted sampling for centrality analysis and community detection in attributed networks. IEEE Transactions on Network Science and Engineering **6**(4), 684–697 (2019)
3. Lee, D.J.L., Han, J., Chambourova, D., Kumar, R.: Identifying fashion accounts in social networks. In: In Proceedings of the KDD Workshop on ML Meets Fashion (2017)
4. Baroni, A., Conte, A., Patrignani, M., Ruggieri, S.: Efficiently clustering very large attributed graphs. In: 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 369–376 (2017)
5. Berahmand, K., Mohammadi, M., Saberi-Movahed, F., Li, Y., Xu, Y.: Graph regularized nonnegative matrix factorization for community detection in attributed networks. IEEE Transactions on Network Science and Engineering (2022)
6. Nasiri, E., Berahmand, K., Li, Y.: Robust graph regularization nonnegative matrix factorization for link prediction in attributed networks. Multimedia Tools and Applications **82**(3), 3745–3768 (2023)
7. Shrum, W., Cheek Jr, N.H., MacD, S.: Friendship in school: Gender and racial homophily. Sociology of Education, 227–239 (1988)
8. McPherson, M., Smith-Lovin, L., Cook, J.M.: Birds of a feather: Homophily in social networks. Annual review of sociology **27**(1), 415–444 (2001)
9. Mislove, A., Viswanath, B., Gummadi, K.P., Druschel, P.: You are who you know: inferring user profiles in online social networks. In: Proceedings of the Third ACM International Conference on Web Search and Data Mining, pp. 251–260 (2010)
10. Espín-Noboa, L., Karimi, F., Ribeiro, B., Lerman, K., Wagner, C.: Explaining classification performance and bias via network structure and sampling technique. Applied Network Science **6**(79) (2021)
11. Wagner, C., Singer, P., Karimi, F., Pfeffer, J., Strohmaier, M.: Sampling from social networks with attributes. In: Proceedings of the 26th International Conference on World Wide Web. WWW '17, pp. 1181–1190, Republic and Canton of Geneva, CHE (2017)
12. Karimi, F., Génois, M., Wagner, C., Singer, P., Strohmaier, M.: Homophily influences ranking of minorities in social networks. Scientific Reports **8**(1), 11077 (2018)
13. Espín-Noboa, L., Wagner, C., Strohmaier, M., Karimi, F.: Inequality and inequity in network-based ranking and recommendation algorithms. Scientific Reports **12**(1), 2012 (2022)
14. Jordan, J.: Geometric preferential attachment in non-uniform metric spaces. Electronic Journal of Probability **18**, 1–15 (2013)
15. Antunes, N., Banerjee, S., Bhamidi, S., Pipiras, V.: Attribute network models, stochastic approximation, and network sampling and ranking. Preprint arXiv:2304.08565v1 (2023)
16. Antunes, N., Bhamidi, S., Guo, T., Pipiras, V., Wang, B.: Sampling based estimation of in-degree distribution for directed complex networks. Journal of Computational and Graphical Statistics **30**(4), 863–876 (2021)
17. Antunes, N., Guo, T., Pipiras, V.: Sampling methods and estimation of triangle count distributions in large networks. Network Science **9**, 134–156 (2021)
18. Grover, A., Leskovec, J.: Node2vec: Scalable feature learning for networks. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '16, pp. 855–864. Association for Computing Machinery, New York, NY, USA (2016)
19. Antunes, N., Bhamidi, S., Pipiras, V.: Learning attribute distributions through random walks. In: Cherifi, H., Mantegna, R.N., Rocha, L.M., Cherifi, C., Micciche, S. (eds.) Complex Networks and Their Applications XI, pp. 17–29. Springer, Cham (2023)
20. Barabási, A.-L., Albert, R.: Emergence of scaling in random networks. science **286**(5439), 509–512 (1999)
21. Krapivsky, P.L., Redner, S.: Organization of growing random networks. Physical Review E **63**(6), 066123 (2001)

22. Bianconi, G., Barabási, A.-L.: Bose-Einstein condensation in complex networks. Physical review letters **86**(24), 5632 (2001)
23. de Almeida, M.L., Mendes, G.A., Madras Viswanathan, G., da Silva, L.R.: Scale-free homophilic network. The European Physical Journal B **86**(2), 38 (2013)
24. Flaxman, A.D., Frieze, A.M., Vera, J.: A geometric preferential attachment model of networks ii. Internet Mathematics **4**(1), 87–111 (2007)
25. Park, J., Barabási, A.-L.: Distribution of node characteristics in complex networks. Proceedings of the National Academy of Sciences **104**(46), 17916–17920 (2007)
26. Meng, L., Masuda, N.: Analysis of node2vec random walks on networks. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences **476**(2243), 20200447 (2020)
27. Kolaczyk, E.D.: Statistical Analysis of Network Data. Springer, New York (2009)
28. Aldous, D., Fill, J.A.: Reversible Markov Chains and Random Walks on Graphs. Unfinished monograph, recompiled 2014, available at http://www.stat.berkeley.edu/ aldous/RWG/book.html (2002)
29. Kumar, S., Sundaram, H.: Attribute-guided network sampling mechanisms. ACM Transactions on Knowledge Discovery from Data **15**(4) (2021)
30. Alon, N., Benjamini, I., Lubetzky, E., Sodin, S.: Non-backtracking random walks mix faster. Communications in Contemporary Mathematics **09**(04), 585–603 (2007)
31. Mihail, M., Papadimitriou, C., Saberi, A.: On certain connectivity properties of the internet topology. In: 44th Annual IEEE Symposium on Foundations of Computer Science, 2003. Proceedings., pp. 28–35 (2003). IEEE
32. Ben-Hamou, A., Lubetzky, E., Peres, Y.: Comparing mixing times on sparse random graphs. In: Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 1734–1740 (2018). SIAM
33. Berger, N., Borgs, C., Chayes, J.T., Saberi, A.: Asymptotic behavior and distributional limits of preferential attachment graphs. The Annals of Probability **42**(1), 1–40 (2014)
34. Garavaglia, A., Hazra, R.S., van der Hofstad, R., Ray, R.: Universality of the local limit of preferential attachment models. arXiv preprint arXiv:2212.05551 (2022)
35. Banerjee, S., Deka, P., Olvera-Cravioto, M.: Local weak limits for collapsed branching processes with random out-degrees. arXiv preprint arXiv:2302.00562 (2023)
36. Chen, N., Litvak, N., Olvera-Cravioto, M.: Generalized pagerank on directed configuration networks. Random Structures & Algorithms **51**(2), 237–274 (2017)
37. Courtney, O.T., Bianconi, G.: Weighted growing simplicial complexes. Physical Review E **95**(6), 062301 (2017)
38. Battiston, F., Cencetti, G., Iacopini, I., Latora, V., Lucas, M., Patania, A., Young, J.-G., Petri, G.: Networks beyond pairwise interactions: structure and dynamics. Physics Reports **874**, 1–92 (2020)
39. Majhi, S., Perc, M., Ghosh, D.: Dynamics on higher-order networks: A review. Journal of the Royal Society Interface **19**(188), 20220043 (2022)
40. Iacopini, I., Petri, G., Barrat, A., Latora, V.: Simplicial models of social contagion. Nature communications **10**(1), 2485 (2019)
41. Fan, J., Yin, Q., Xia, C., Perc, M.: Epidemics on multilayer simplicial complexes. Proceedings of the Royal Society A **478**(2261), 20220059 (2022)
42. Sun, H., Radicchi, F., Kurths, J., Bianconi, G.: The dynamic nature of percolation on networks with triadic interactions. Nature Communications **14**(1), 1308 (2023)