Learning Attribute Distributions Through Random Walks

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

[1] Center for Computational and Stochastic Mathematics, University of Lisbon and University of Algarve, Avenida Rovisco Pais 1049-001, Lisbon, Portugal
nantunes@ualg.pt,

[2] Department of Statistics and Operations Research, University of North Carolina, CB 3260, Chapel Hill, NC 27599, USA
bhamidi@email.unc.edu, pipiras@email.unc.edu

Abstract. We investigate the statistical learning of nodal attribute distributions 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, where new nodes form connections dependent on both their attribute values and popularity as measured by degree. 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. The performance of such algorithms is studied on both synthetic networks and real world systems, and its dependence on the degree of homophily, or absence thereof, is assessed.

Keywords: Attributed networks, homophily, network model, random walk samplings, discrete and continuous attributes, learning distributions.

1 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. In one direction, machine learning pipelines such as network representation learning [10], clustering [8], classification [17], and community detection [6] have been developed to study the entire network. Driven by the scale of data, the main motivation of this paper, is network sampling, where limited explorations are used to learn network level functionals such as the degree distribution [19].

One standard phenomenon in many such real world systems is *homophily* [22, 18, 20], i.e., node pairs with similar attributes being likelier connected than node pairs with discordant attributes. 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 [23]; the effect of homophily on centrality measures and visibility of minority groups and fairness questions [14]. This paper studies the estimation of the attribute distribution (both discrete and continuous) for homophily networks. We extend the attributed driven preferential attachment model [14, 13] 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. Uniform random sampling of nodes or edges is the "gold standard", providing unbiased estimates of corresponding attribute distributions. However, owing to both computational and privacy issues in settings such as social networks, such sampling is often infeasible. In these cases, link trace sampling, such as random walks (RW) are typically used; see references in [3, 4] for estimation of functionals such as degree distribution and clustering. Much less is known in the context of attribute distribution estimation. In this paper, we consider several canonical attribute agnostic sampling schemes such as Metropolis-Hasting random walk, versions of node2vec [12] that incorporate both classical random walk and non-backtracking propensities and propose variants of node2vec where edge weights depend on attributes of the node pair. The performance of the considered random walk sampling schemes in terms of estimation error of the attribute distributions is studied across the following four dimensions in both synthetic and real world settings: **(a)** Inherent homophilc 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 the network.

Overview of findings and organization of the paper: 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 seems to improve performance, especially in conjunction with attribute dependent weights; (iv) the performance of RWs is well below the "gold standard" of random node sampling; (v) methods seem to work comparably well for discrete and continuous attributes.

The paper is organized as follows. A synthetic model with homophily is given in Sec. 2. Sampling schemes for learning attribute distribution are described in Sec. 3. Statistical learning tasks are discussed in Sec. 4. Numerical evaluation on synthetic and real data are described in Sec. 5. Sec. 6 concludes.

2 Attribute Network Models with Homophily

We now describe the main synthetic model, termed non-linear preferential attachment (NLPA) model with homophily. 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 $\alpha \geq 0$ playing the role of degree in measuring popularity. Let N be the number of nodes (vertices) in the network. Nodes $\{v_t : 1 \leq t \leq N\}$ enter the system sequentially starting at $t = 1$ with a base connected graph $\mathcal{G}_1$ with every node having an attribute in $\mathcal{A}$. Every node v_t has attribute $a(v_t) \in \mathcal{A}$ generated independently using μ. The dynamics are recursively defined as follows: for any t and $v \in \mathcal{G}_t$, let $\deg(v,t)$ denote the degree of v at time t. Conditional on $\mathcal{G}_t$, the probability that v_{t+1} connects to $v \in \mathcal{G}_t$ is proportional to:

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

The model (1) extends various existing models including: Barabási-Albert model [5] ($f \equiv 1$, $\alpha = 1$), sublinear PA [16] ($f \equiv 1$, $0 < \alpha < 1$), PA with multiplicative fitness [7] ($f(a,a') = a$, $\alpha = 1$), scale free homophilic model [9] ($f(a,a') = 1 - |a - a'|$, $\mathcal{A} = [0,1]$, $\alpha = 1$), and geometric versions with $\alpha = 1$, $\mathcal{A}$ a compact metric space and f an appropriate function of the distance [11, 13]. Most existing studies focus on asymptotics for either the degree distribution or maximal degree.

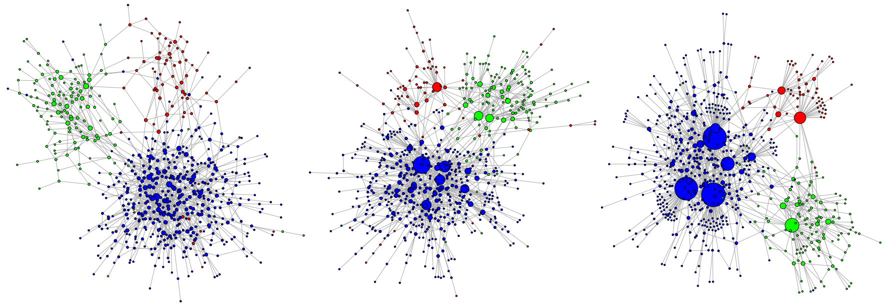
When the latent space $\mathcal{A} = \{1, 2, \ldots, K\}$ is finite, one can define, macroscopic measures of homophily, and the converse heterophily from an observed network $\mathcal{G}$ (either synthetic or empirically observed) on N nodes as follows [21]. Let $\mathcal{E}$ denote the total edge set; for $a \in \mathcal{A}$, $\mathcal{V}_a$ 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 type a and a'. Let $p = |\mathcal{E}|/\binom{N}{2}$ be the edge density. For $a \in \mathcal{A}$, $D_a = |\mathcal{E}_{aa}|/(\binom{|\mathcal{V}_a|}{2}p)$ 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. Similarly, for $a \neq a'$, $H_{aa'} = |\mathcal{E}_{aa'}|/(|\mathcal{V}_a||\mathcal{V}_{a'}|p)$ denotes propensity of type a nodes to connect to type a' nodes as contrasted with random placement of edges at the same level as the global edge density.

An illustration of synthetic networks generated using the NLPA model (1) with finite latent space is given in Fig. 1. Here, $\mathcal{A} = \{1, 2, 3\}$ represent 70%, 20% and 10% of the total $N = 1000$ nodes, resp.; $f(a,a) = 0.95$, $f(a,a') = 0.025$, for $a \neq a' = 1, 2, 3$. The network is plotted for different values of α – Fig. 1(a)–1(c). For $\alpha = 0.2$, the corresponding homophily measures are $D_1 = 1.45$, $D_2 = 4.36$, $D_3 = 7.38$, $H_{12} = 0.07$, $H_{13} = 0.14$, $H_{23} = 0.45$. For $\alpha = 1.2$, the homophily measures are $D_1 = 1.38$, $D_2 = 4.84$, $D_3 = 9.12$, $H_{12} = 0.08$, $H_{13} = 0.08$, $H_{23} = 0.16$.

3 Network Sampling Schemes

This section describes sampling schemes for learning attribute distribution, both random walk based, as well as corresponding "gold standard" schemes. Throughout this section, for graph $\mathcal{G}$ and node $i \in \mathcal{G}$, d_i will denote its degree.

Metropolis Hasting 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 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



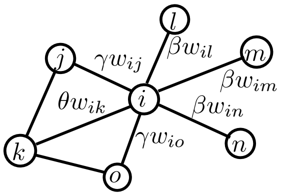
(a) (b) (c)

Fig. 1. Networks generated by the NLPA model with (a) $\alpha = 0.2$, (b) $\alpha = 1$, (c) $\alpha = 1.2$.

nodes. It is easy to check that the stationary distribution is uniform over the node set, i.e., $\pi_i = 1/N$ for $1 \leq i \leq N$.

Node2vec (N2V). As proposed in [12], 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 previous and current visited nodes be k and i, resp. The next visited node j is chosen according to the transition probability proportional to:

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



where w_{ij} is the weight of edge (i,j) - see figure. We now describe specific variants of this class of random walks.

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

$$\pi_i = \frac{d_i}{2|\mathcal{E}|}. \tag{2}$$

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. The stationary distribution in this case is given by

$$\pi_i \propto \sum_j w_{ij}. \tag{3}$$

Node2vec-3 (N2V-3): If the network is simple (i.e. unweighted, undirected, without self-loops and multiple edges) and $\beta = \gamma$, $\theta > 0$, the stationary distribution for nodes is given by Eq. (2). With small θ, the walk approaches the non-backtracking random walk.

Node2vec-4 (N2V-4): One can consider other variants of N2V. We consider below 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 distribution. Analogous to the stationary distribution for N2V-3 matching the usual RW in the stationary regime, it is expected that especially in the small θ setting, the stationary distribution can still be approximated by that in Eq. (3). We explore the efficacy of this approximation for moderate size synthetic networks below.

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 (and their attributes) 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. Since ES selects edges rather than nodes to populate the sample, the node (attribute) set is constructed by including both incident nodes (attributes) in the sample when a particular edge is sampled.

4 Statistical Learning Methods

We now discuss the estimation of attribute distributions from the data collected through RWs, with discrete attributes described in Sec. 4.1 and continuous attributes in Sec. 4.2.

4.1 Discrete Attributes

Run a random walk (any of the schemes described in Sec. 3) for n steps and let i_s denote the s-th node sampled by a RW, for $1 \leq s \leq n$. Since nodes are sampled with replacement and with probabilities π_i in the stationary regime, the attribute distribution can be estimated as

$$\hat{p}(a) = \frac{1}{Nn} \sum_{s=1}^{n} \frac{\mathbf{1}\{a(i_s) = a\}}{\pi_{i_s}}, \qquad a \in \mathcal{A}, \tag{4}$$

where $\mathbf{1}\{B\} = 1$ if B is true and 0 otherwise [15] (Chapter 5). If the total number of nodes N is unknown, its estimator is given by $(1/n) \sum_s 1/\pi_{i_s}$. For N2V-2 this results in,

$$\hat{p}(a) = \frac{1}{\sum_{s=1}^{n} 1/w_{i_s}} \sum_{s=1}^{n} \frac{\mathbf{1}\{a(i_s) = a\}}{w_{i_s}}, \qquad a \in \mathcal{A}. \tag{5}$$

For fixed a, the MSE of $\hat{p}(a)$ is given by $E[(\hat{p}(a) - p(a))^2]$. In the stationary regime, $\hat{p}(a)$ in (4) is an unbiased estimator of $p(a)$ and the MSE is equal to the variance $V[\hat{p}(a)]$. The variance of $\hat{p}(a)$ 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 [1] (Proposition 4.29), we have

$$V(\hat{p}(a)) \leq \frac{2\Lambda(a)}{\delta n}\left(1 + \frac{\delta}{2n}\right), \tag{6}$$

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 attribute a is upper bounded by the inverse of the spectral gap and $\Lambda(a)$, the latter is small if the probability of sampling nodes with attribute a is large. We will see in Sec. 5 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.

4.2 Continuous Attributes

Let $g(\cdot)$ be the density of the continuous attributes, and as before $(i_s : 1 \leq s \leq n)$ be the states visited by the RW with corresponding attributes $(a(i_s) : 1 \leq s \leq n)$. Analogous to (5) the natural estimate for $g(\cdot)$ is through standard kernel smoothing as

$$\hat{g}(a) = \sum_{s=1}^{n} K\left(\frac{a - a(i_s)}{h}\right)\frac{1}{h}w_s, \tag{7}$$

where $h > 0$ is a bandwidth, K is a kernel function, and the weights w_s satisfy

$$w_s \propto \frac{1}{\pi_{i_s}}, \quad \sum_{s=1}^{n} w_s = 1. \tag{8}$$

The performance of the estimator can be assessed through the estimation error: for $q > 0$,

$$\mathit{error} = \left[\int |\hat{g}(a) - g(a)|^q da\right]^{1/q}. \tag{9}$$

The values of q usually considered are 1 and 2.

5 Numerical Studies

5.1 Synthetic Networks

We consider the NLPA model in (1) for networks with attributes and explore the effect of homophily on the accuracy of the RWs to estimate the attribute distribution in a controlled setting.

Random walks	MH	N2V-1	N2V-2$^{\mathrm{a}}$	N2V-2$^{\mathrm{b}}$	N2V-2$^{\mathrm{c}}$	N2V-3	N2V-4	NS	ES
st. dev.; $a = 1$	0.235	0.172	0.199	0.142	0.169	0.123	0.102	0.029	0.051
st. dev.; $a = 2$	0.198	0.152	0.176	0.127	0.151	0.107	0.089	0.025	0.042
st. dev.; $a = 3$	0.137	0.077	0.098	0.069	0.086	0.065	0.049	0.018	0.035
spectral gap (δ)	0.019	0.048	0.037	0.040	0.011	0.107	0.106	-	-
$\Lambda(3)$	0.088	0.149	0.165	0.135	0.243	0.149	0.135	-	-

Table 1. Standard deviations, spectral gaps and quantities $\Lambda(3)$, under the PA model with $\alpha = 0.2$ and attribute values $a = 1, 2, 3$. For the RW weights: N2V-2$^{\mathrm{a}}$ ($\overline{w}_{aa} = 1.5, \overline{w}_{aa'} = 1$), N2V-2$^{\mathrm{b}}$ ($\overline{w}_{aa} = 0.3, \overline{w}_{aa'} = 1$), N2V-2$^{\mathrm{c}}$ ($\overline{w}_{aa} = 0.05, \overline{w}_{aa'} = 1$).

Discrete Attributes. The total number of nodes is $N = 2,000$ with attributes labeled $a = 1, 2, 3$. If a node attribute is selected at random, its p.m.f. is given by $p(1) = 0.7$, $p(2) = 0.2$ and $p(3) = 0.1$. The tendency of two nodes to connect according to the NLPA model is $f(a, a) = 0.9$, $f(a, a') = 0.05$, $a, a' = 1, 2, 3$, $a \neq a'$. Consider first the case $\alpha = 0.2$, where the number of nodes with a large degree tends to be smaller – see Fig. 1(a). For the largest component of the generated network, the homophily measures are $D_1 = 1.39$, $D_2 = 3.93$, $D_3 = 7.26$, $H_{12} = 0.17$, $H_{13} = 0.10$, $H_{23} = 0.35$.

The network attributes are sampled with the different RWs on the largest component and the p.m.f. of the attributes is estimated using (4). Table 1 shows the standard deviations of the estimates using 300 runs for each RW with length $0.15N$. The MH walk presents the worst performance. Compared to the baseline method NS that samples nodes according to the limit stationary distribution of MH, the diference in variability is large. The N2V-1 walk performs the worst among the variants of N2V. It represents the classical RW since edges are sampled at random in its stationary limit. However, the variability of the baseline method ES is smaller. The results for MH and N2V-1 can also be explained through the bound of the variance (6). The spectral gap δ is sufficiently larger for N2V-1, resulting in a lower variability for attribute $a = 3$, in spite of smaller $\Lambda(3)$ for MH.

We examine how the different choices of weights affect the performance of N2V-2. We write $\overline{w}_{aa}$ for the weights of nodes with the same attributes, and $\overline{w}_{aa'}$ with different attributes. If $\overline{w}_{aa}$ is greater than $\overline{w}_{aa'}$ (N2V-2$^{\mathrm{a}}$ in Table 1), the RW hardly transits from one attribute value to another, which creates a bottleneck for approaching the stationary probability. On the other hand, if $\overline{w}_{aa}$ is smaller than $\overline{w}_{aa'}$ (N2V-2$^{\mathrm{b}}$), movements between different attribute values are more frequent, accelerating the convergence. In this case, the spectral gap increases. However, as the difference between $\overline{w}_{aa'}$ and $\overline{w}_{aa}$ incresases (N2V-2$^{\mathrm{c}}$), the convergence is decelerated because exploration within the same attribute is not sufficient due to the inter-attribute moves. We also see that if $\overline{w}_{aa'}$ is greater than $\overline{w}_{aa}$ until a certain point, the probability of the random walker of sampling nodes with attribute $a = 3$ increases and $\Lambda(3)$ decreases (see the discussion below (6)). The tradeoff between δ and $\Lambda(a)$ explains the smaller variability for the three attribute values of N2V-2$^{\mathrm{b}}$, which outperforms N2V-1.

	Random walks	MH	N2V-1	N2V-2	N2V-3	N2V-4	NS	ES
with homophily	st. dev.; $a = 1$	0.291	0.153	0.131	0.126	0.100	0.031	0.054
	st. dev.; $a = 2$	0.256	0.131	0.107	0.108	0.090	0.028	0.045
	st. dev.; $a = 3$	0.160	0.094	0.073	0.068	0.059	0.021	0.036
without homophily	st. dev.; $a = 1$	0.146	0.059	0.055	0.049	0.045	0.029	0.040
	st. dev.; $a = 2$	0.116	0.055	0.051	0.041	0.039	0.025	0.036
	st. dev.; $a = 3$	0.110	0.039	0.036	0.031	0.024	0.018	0.027

Table 2. Standard deviations for various RWs (N2V-2 with $\overline{w}_{aa} = 0.3, \overline{w}_{aa'} = 1$), under the NLPA model with $\alpha = 1$ and with/without homophily and attribute values $a = 1, 2, 3$.

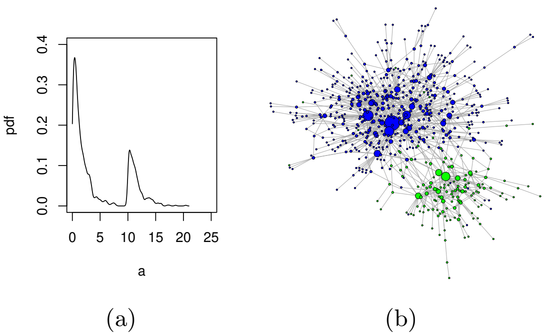
In N2V-3, the parameter θ of the propensity for the random walk to backtrack is decreased to $\theta = 10^{-3}$ and $\beta = \gamma = 1$ are kept for the other two parameters. (Note that if the walker arrives at a node with degree 1, it always backtracks in the next time step since this is the only possible move.) In this case, a random walker tends to explore better the network within the same attribute value, which accelerates the convergence. The result is consistent with the non-backtracking RWs on regular graphs [2]. 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 weighted and non-backtracking RWs. We use the same weights and backtracking parameter as in N2V-2[b] and N2V-3, resp. Since the stationary distribution is not known, we approximate it using (3). The choice is heuristic but the results show that N2V-4 has lower variability. This can be explained by the decrease of $\Lambda(a)$ for attribute values 2 and 3 (see $\Lambda(3)$ for N2V-3 and N2V-4 while δ is approximately equal). We have confirmed these findings by using the true stationary distribution of N2V-4 obtained through simulation.

We next consider the NLPA network with $\alpha = 1$ and take its remaining parameters as above. For the largest component of the network, the homophily measures are $D_1 = 1.38$, $D_2 = 4.30$, $D_3 = 6.25$, $H_{12} = 0.16$, $H_{13} = 0.23$, $H_{23} = 0.32$. The standard deviation of 300 runs for each RW is given in Table 2. In this case, the standard deviation of MH increases and of N2V-1 decresases. This can be explained by nodes with different attribute values attracted to high degree nodes – see Fig. 1(b). Unlike the case $\alpha = 0.2$, the RWs which are attracted by high degree nodes will benefit from this to move between different attribute values. The same conclusions can be drawn as above for the other variants of N2V.

Finally, we consider a network without homophily where f is constant and $\alpha = 1$. The results are shown in Table 2. As seen from the table, if the homopliy decreases, the differences between the RWs tend to be smaller.

Continuous Attributes. We consider the NLPA model with N=2,000 nodes and $\alpha = 1$. Nodes have continuous attributes with values drawn independently from the following probability distribution. Let X be a gamma random variable with shape and scale parameters 1 and 1.5, resp. For the attributes, we draw



(a) (b)

Fig. 2. (a) Probability density function of attributes estimated using kernel smoothing (b) The generated NLPA network with attributes less than 10 (blue) and greater than 10 (green).

$0.7N$ and $0.3N$ independent random variables X and $10 + X$, resp. The density function of attributes estimated using kernel smoothing is shown in Fig. 2(a). Additionally, we set

$$f(a(i), a(j)) = \begin{cases} 0.95, \; a(i), a(j) < 10 \text{ or } a(i), a(j) > 10, \\ 0.05, \text{ otherwise.} \end{cases} \tag{10}$$

The network generated is plotted in Fig. 2(b), where nodes are divided in two groups: with attributes less than 10 (group 1) and greater than 10 (group 2). The homophily measures are $D_1 = 1.378$, $D_2 = 3.092$, $H_{12} = H_{21} = 0.112$.

For N2V-2, the weights are taken as $w_{ij} = |a(i) - a(j)|^b$, which allows moving between the groups of nodes but also giving more weight to edges with different values within each group. 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. From the experiments, we found that values of b close to zero decrease the range of weights and show good results.

The network attributes are sampled with the different sampling methods on the largest component and the density function of the attributes is estimated using (7). Table 3 shows the average of the estimation error (9) with $q = 1$ and the spectral gap from 300 runs for each RW with length $0.15N$. We fixed $b = 0.3$ (N2V-2/4) and $\theta = 10^{-3}$ (N2V-3/4). The performance of the samplings methods is akin to the case of the discrete attributes.

5.2 Real Networks

We analyze two publicly available datasets of real networks with attributes and homophily.[3]

[3] https://snap.stanford.edu/data/

Random walks	MH	N2V-1	N2V-2	N2V-3	N2V-4	NS	ES
average error	0.818	0.487	0.457	0.385	0.364	0.186	0.270
spectral gap (δ)	0.005	0.042	0.051	0.080	0.096	-	-

Table 3. Estimation error and spectral gap for various RWs, under the NLPA model.

Random walks	N2V-1	N2V-2	N2V-3
politician (1)	0.052	0.0561	0.0489
government (2)	0.051	0.046	0.043
tv show (3)	0.047	0.045	0.038
company (4)	0.0669	0.058	0.054

Table 4. St. dev. of the estimates.

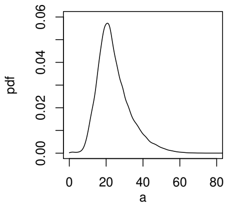


Fig. 3. P.d.f. of the age.

Discrete Attributes. The dataset is a webgraph of Facebook sites. Nodes represent pages while the links are mutual likes between sites. Node features were extracted from the site descriptions that the page owners created to summarize the purposes of the sites. The graph was collected through the Facebook Graph API and restricted to pages from four attributes which are defined by Facebook. These attributes are: politicians (1), governmental organizations (2), television shows (3) and companies (4). We consider the simplified network which has $N = 22,470$ nodes and $170,823$ edges. The distribution of node attributes is $p(1) = 0.31$, $p(2) = 0.26$, $p(3) = 0.29$, and $p(4) = 0.14$. The homophily measures are $D_1 = 3.28$, $H_{1*} = 0.17$, $D_2 = 5.08$, $H_{2*} = 0.21$, $D_3 = 3.46$, $H_{3*} = 0.14$, $D_4 = 1.41$, $H_{4*} = 0.11$, where H_{a*} denotes the propensity of attribute a nodes to connect to the other types of attributes.

To estimate the p.m.f. of the node attributes, we consider only the variants of N2V with known stationary distributions. For N2V-2, we set the weights as $\overline{w}_{aa} = 0.3$, $\overline{w}_{aa'} = 1$, $a, a' = 1, 2, 3, 4$, $a \neq a'$, and for N2V-3, we set $\theta = 10^{-3}$. Table 4 shows the standard deviations of the estimates using RWs of length $0.15N$ and 500 runs. The results are in line with the synthetic model with discrete attributes where sampling with N2V-3 produces more accurate estimates.

Continuous Attributes. Pokec is a social network with attributes from Slovakia. We use the age attribute viewed as continuous as in [24]. Considering only the nodes with age attributes results in a network with N=1,138,314 nodes and 22,301,601 edges - see Fig. 3. It is well known that the network is moderately homophilic with respect to age. If we divide the nodes in two groups: say, age less or equal to 37 (group 1) and greater than 37 (group 2), the homophilic measures of the groups are $D_1 = 1.166$, $H_{12} = 0.30$ and $D_2 = 1.46$. Group 2 represents 9% of the total number of nodes. The average of the estimation errors from 20 runs for each RW with length $0.05N$ are: 0.036 (N2V-1), 0.031 (N2V-2 with $b = 0.2$) 0.030 (N2V-3 with $\theta = 10^{-3}$).

6 Discussion and Future Directions

In this paper, we developed a statistical learning framework for the attribute distributions in networks and evaluated numerically the impact of homophily, degree centrality, and random walk exploration mechanisms on estimation accuracy. The results seem to indicate intricate non-linear relationship between intrinsic homophilic characteristics of the network, parameters modulating random walk exploration schemes and the error of proposed learning algorithms. Untangling the precise relationship will require careful theoretical understanding both of macroscopic functionals such as the spectral gap of proposed RWs and their relationship to parameters such as backtracking propensities and jump rates across different attribute sets, as well as microscopic functionals such as asymptotics for local neighborhoods of the underlying network. This should lead to more principled ways of choosing RWs and their parameters in terms of the network homophily, centrality and possibly other measures.

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 [14]. Other questions, including learning joint distributions of the degree and the attribute through sampling mechanisms, as well as multivariate attribute distributions, both in terms of developing synthetic models, as well as real world data will also be considered.

Acknowledgements. The second and third authors were supported in part by the NSF grants DMS-2113662 and DMS-2134107.

References

1. D. Aldous and J. A. Fill. Reversible Markov chains and random walks on graphs, 2002. Unfinished monograph, recompiled 2014, available at http://www.stat.berkeley.edu/~aldous/RWG/book.html.
2. N. Alon, I. Benjamini, E. Lubetzky, and S. Sodin. Non-backtracking random walks mix faster. *Communications in Contemporary Mathematics*, 09(04):585–603, 2007.
3. N. Antunes, S. Bhamidi, T. Guo, V. Pipiras, and B. Wang. Sampling based estimation of in-degree distribution for directed complex networks. *Journal of Computational and Graphical Statistics*, 30(4):863–876, 2021.
4. N. Antunes, T. Guo, and V. Pipiras. Sampling methods and estimation of triangle count distributions in large networks. *Network Science*, 9(S1):S134–S156, 2021.
5. A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.
6. A. Baroni, A. Conte, M. Patrignani, and S. Ruggieri. Efficiently clustering very large attributed graphs. In *2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 369–376, 2017.
7. G. Bianconi and A.-L. Barabási. Bose-einstein condensation in complex networks. *Physical review letters*, 86(24):5632, 2001.

8. C.-H. Chang, C.-S. Chang, C.-T. Chang, D.-S. Lee, and P.-E. Lu. Exponentially twisted sampling for centrality analysis and community detection in attributed networks. *IEEE Transactions on Network Science and Engineering*, 6(4):684–697, 2019.
9. M. L. de Almeida, G. A. Mendes, G. Madras Viswanathan, and L. R. da Silva. Scale-free homophilic network. *The European Physical Journal B*, 86(2):38, 2013.
10. H. Fan, Y. Zhong, G. Zeng, and L. Sun. Attributed network representation learning via improved graph attention with robust negative sampling. *Applied Intelligence*, 51(1):416–426, 2021.
11. A. D. Flaxman, A. M. Frieze, and J. Vera. A geometric preferential attachment model of networks ii. *Internet Mathematics*, 4(1):87–111, 2007.
12. A. Grover and J. Leskovec. Node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 855–864, New York, NY, USA, 2016. Association for Computing Machinery.
13. J. Jordan. Geometric preferential attachment in non-uniform metric spaces. *Electronic Journal of Probability*, 18:1–15, 2013.
14. F. Karimi, M. Génois, C. Wagner, P. Singer, and M. Strohmaier. Homophily influences ranking of minorities in social networks. *Scientific Reports*, 8(1):11077, 2018.
15. E. D. Kolaczyk. *Statistical Analysis of Network Data*. Springer Series in Statistics, 2009.
16. P. L. Krapivsky and S. Redner. Organization of growing random networks. *Physical Review E*, 63(6):066123, 2001.
17. D. J. L. Lee, J. Han, D. Chambourova, and R. Kumar. Identifying fashion accounts in social networks. In *In Proceedings of the KDD Workshop on ML Meets Fashion*, 2017.
18. M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444, 2001.
19. L. Meng and N. Masuda. Analysis of node2vec random walks on networks. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 476(2243):20200447, 2020.
20. A. Mislove, B. Viswanath, K. P. Gummadi, and P. Druschel. 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*, pages 251–260, 2010.
21. J. Park and A.-L. Barabási. Distribution of node characteristics in complex networks. *Proceedings of the National Academy of Sciences*, 104(46):17916–17920, 2007.
22. W. Shrum, N. H. Cheek Jr, and S. MacD. Friendship in school: Gender and racial homophily. *Sociology of Education*, pages 227–239, 1988.
23. C. Wagner, P. Singer, F. Karimi, J. Pfeffer, and M. Strohmaier. Sampling from social networks with attributes. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, page 1181–1190, Republic and Canton of Geneva, CHE, 2017.
24. H. Yang, W. Xiong, X. Zhang, K. Wang, and M. Tian. Penalized homophily latent space models for directed scale-free networks. *PLoS One*, 16(8):e0253873, 2021.