Minority representation and relative ranking in sampling attributed networks

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

[1] Center for Computational and Stochastic Mathematics, University of Lisbon, 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
sayan@email.unc.edu, bhamidi@email.unc.edu, pipiras@email.unc.edu

Abstract. We explore two questions related to sampling and minorities in attributed networks with homophily. The first question is to investigate sampling schemes which favor minority attribute nodes and which give preference to "more popular" nodes having higher centrality measures in the network. A data study shows the efficiency of Page-rank and walk-based network sampling schemes on a directed network model and a real-world network with small minorities. The second question concerns the effect of homophily and out-degrees of nodes on the relative ranking of minorities compared to majorities in degree-based sampling. Several synthetic network configurations are considered and the conditions for minority nodes to have a higher relative rank are investigated numerically. The results are also assessed with real-world networks.

Keywords: Random networks, attributes, homophily, sampling, minorities, ranking.

1 Introduction

An attributed network can be defined as a graph in which nodes (and/or edges) have features. In a social network, node attributes can refer to gender, age, ethnicity, political ideologies. The attributes of nodes often co-vary and affect the graph structure. One standard phenomenon in many real-world systems is homophily [6], 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 an attribute. Additionally, the distribution of user attributes over the network is usually uneven, with coexisting groups of different sizes, e.g., one ethnic group (majority) may dominate other (minority).

Given that most real networks can only be observed indirectly, network sampling, and its impact on the representation/learning of the true network, is an activate area of research across multiple communities (see e.g. [2, 3] and the references therein). In this context, there has been significant interest in attributed network sampling where there is a particular *small minority* of certain attribute

nodes. Here, we explore two related questions in this area, namely, (a) settings where Page-rank and other exploration based sampling schemes favor sampling small minorities, (b) effects of homophily and out-degrees on the relative ranking of minorities compared to majorities in degree-based sampling. To this end, we shall use an attributed network model that incorporates homophily [1]. We employ the asymptotic theory developed in [1] to gain insight through data studies of the various network sampling schemes and attribute representation in concrete applications. The findings will also be assessed with real-world networks. More concretely, we investigate the following research problems:

(a) We consider the case where there is a particular small minority which has higher propensity to connect within itself as opposed to majority nodes; for substantial recent applications and impact of such questions, see [11, 10, 13]. In such setting, devising schemes where one gets a non-trivial representation of minorities is challenging if the sample size is much smaller than the network size. In this case, uniform sampling will clearly not be *fair* as the sampled nodes will tend to be more often from the majority attribute. Additionally, uniform sampling does not give preference to "more popular" minority nodes, i.e., higher degree/Page-rank nodes. Therefore, it is desirable to *explore* the network locally around the initial (uniformly sampled) random node and try to travel towards the "centre", thereby traversing edges along their natural direction. However, to avoid high sampling costs, the explored set of nodes should not be too large. We compare through a data study several sampling schemes derived from centrality measures like degree and Page-rank and show that they increase the probability of sampling a minority node and its "popularity". This is investigated in several network model configurations and in a real network dataset.

(b) We consider two degree-based sampling schemes and explore the effects of homophily and out-degrees of the model parameters on the relative ranking of minority compared to majority (in terms of proportion) in the samples. As in (a), we again study minority representation, but focus on degree-based sampling and are interested in dependence on structural network properties. The conditions in a asymptotic regime (when the number of nodes goes to infinity) are known for the minority nodes to rank higher (i.e. have larger proportions) than the majority nodes (based on the tail distribution and sum of the degrees) [1]. For three scenarios - heterophily, homogeneous homophily (homogenous mixing) and asymmetric homophily - the results are numerically investigated for the minority nodes to rank higher. The last two scenarios were briefly considered heuristically in [9, 7] using fluid limits. We show that the results for two real networks with degree power-law distributions agree with those for the synthetic model.

The paper is organized as follows. A synthetic model with homophily is given in Sec. 2. Network sampling in the presence of a small minority is studied in Sec. 3. Relative ranking of minorities is investigated in Sec. 4. Sec. 5 concludes and indicates future work.

2 Network Model with Attributes and Homophily

Fix an attribute space $S = \{1, 2\}$. The nodes with attribute 1 will be referred as *minority* and attribute 2 as *majority*. While this paper only deals with these two

types, the setting below can be extended to more general attribute spaces. Fix a probability mass function (π_1, π_2) on S and a possible asymmetric function $\kappa : S \times S \to \mathbb{R}$; this function measures propensities of pairs of nodes to connect, based on their attributes. Fix a preferential attachment parameter $\alpha \in [0, 1]$ and an out-degree function $m : S \to \mathbb{N}$ which modulates the number of edges that a node entering the system connects to, depending on its attribute type.

Nodes enter the system sequentially at discrete times $n \geq 1$ starting with a base connected graph G_0 with n_0 nodes at time $n = 0$ where every node has an attribute in S. Write v_n for the node that enters at time n and $a(v_n)$ for the corresponding attribute; every node v_n has attribute 1 with probability π_1 and attribute 2 with probability π_2. The dynamics of construction are recursively defined as: for $n \geq 0$ and $v \in G_n$, v_{n+1} attaches to the network via $m(a(v_{n+1})) = m_{a(v_{n+1})}$ outgoing edges. Each edge independently chooses an existing node in G_n to attach to, with probabilities (conditionally on G_n and $a(v_{n+1})$) given by

$$\mathbb{P}\left(v_{n+1} \rightsquigarrow v \,|\, G_n, a(v_{n+1}) = a^\star\right) := \frac{\kappa(a(v), a^\star)[\deg(v, n)]^\alpha}{\sum_{v' \in G_n} \kappa(a(v'), a)[\deg(v', n)]^\alpha}, \qquad (1)$$

where $\deg(v, n)$ denote the degree of v at time n (if $G_0 = v_0$, initialize $\deg(v_0, 0) = 1$). A *tree* network is obtained if $m_1 = m_2 = 1$. The case $\kappa(.,.) = 1$ and $\alpha = 1$ corresponds to the well known *linear* preferential attachment model while $0 < \alpha < 1$ to the *sublinear* case. When referring to a synthetic network below, we shall always mean the model (1).

In measuring homophily, we extend the definition given for signed networks [12] to directed networks. Let V (resp. E) denote the set of nodes (resp. edges) of a network; for $a = 1, 2$, let V_a be the set of nodes with attribute a, and for $a \neq a'$, let $E_{aa'}$ be the set of edges between nodes of types a and a'. Let $p = |E|/(|V|(|V| - 1))$ be the edge density. For $a = 1, 2$, dyadicity $D_a = |E_{aa}|/(|V_a|(|V_a|-1)p)$ 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. Similarly, for $a \neq a'$, heterophilicity $H_{aa'} = |E_{aa'}|/(2|V_a||V_{a'}|p)$ 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 type a do not tend to be connected to nodes of type a'.

3 Network Sampling and Minority Representation

In this section, we compare attribute representation of minorities under sampling schemes on synthetic and empirical networks.

3.1 Sampling Methods

We consider several sampling schemes derived from various centrality measures such as (in-)degree and Page-rank. All the methods have in common the following exploration idea of the network. A node is picked uniformly at random followed by a walk on the network with a fixed or random number of steps. The sampled node is the last node visited by the walk.

Uniform sampling (U). We choose a node at random and the number of walk steps is zero. This is equivalent to the classic uniform sampling method.
Sampling proportional to degree (D). We pick a node uniformly at random, and one of its neighbors is chosen at random (one walk step).
Sampling proportional to in-degree (ID). For directed network, a node is selected at random and one step is taken through an out-going edge chosen at random. If the out-degree of the node is zero, the selected node is sampled.
Sampling proportional to Page-rank (PR_c). After choosing an initial node at random, the number of steps to traverse the directed network is a geometric random variable (starting at zero) with parameter $(1-c)$. If the walk gets stuck in a node before the number of steps is reached, it returns this node as the sampled node. The equivalence of this algorithm and sampling proportional to Page-rank with damping factor c in the context of tree network models follows from [5].
Fixed length walk sampling (FL_M). We pick a node uniformly at random and walk a fixed number M of steps through the out-going edges chosen at random of the visited nodes. The same rule above applies if a node with zero out-degree is reached.

3.2 Asymptotic Analysis: Sampling in Tree Networks

We consider an asymmetric homophily scenario. Majority nodes (type 2) have equal propensity to connect to minority (type 1) or majority nodes. Minorities have relatively higher propensity to connect to other minority nodes compared to majority nodes.

Let $\kappa_{11} = \kappa_{22} = \kappa_{12} = 1$, $\kappa_{21} = a$ and $\pi_1 = \theta/(1+\theta)$. We analyze the sampling schemes when a and θ go to zero in a dependent way by letting $\theta = D\sqrt{a}$, where D is a positive constant. Let v be a node sampled from the network G_n and $a(v)$ its attribute, under one of the sampling methods above. From the analysis of the linear, tree model [1], as $a \to 0^+$, we have that $\mathbb{P}(a(v) = 1|G_n)$ behaves as $D\sqrt{a} + O(a)$ under uniform sampling; $2D\sqrt{a} - (4D^2 + \frac{1}{2})a + O(a^{3/2})$ under sampling proportional to degree; $3D\sqrt{a} + O(a)$ under sampling proportional to in-degree; and as $c \to 1^-$ and $n \to \infty$, $(2D^2 - \frac{1}{2} + \Delta)/(2D^2 + \frac{1}{2} + \Delta)$, where $\Delta = \sqrt{(2D^2 - 1/2)^2 + 4D^2}$ under sampling proportional to Page-rank and fixed length walk sampling. The next sections investigate how these results hold in a non-asymptotic regime in (sub-)linear, (non-)tree networks, as well as in a real network.

3.3 Synthetic Networks

We generate a linear ($\alpha = 1$), tree network with $|V| = 10^5$ nodes, $a = 0.003$ ($D = 1$) where the probability that a node entering the network has attribute 1 (minority) is very small, $\pi_1 \approx 0.052$. The homophily and structural characteristics of the network are given in Table 1 (Syn. 1). Note that D_1 is large while D_2 is close to 1, and H_{12} is smaller than H_{21} (< 1) corresponding to an asymmetric homophily. A picture of a small network generated in this setting is shown in

Table 1. Synthetic networks: structural properties.

	$\|V\|$	$\|E\|$	D_1	D_2	H_{12}	H_{21}	$\frac{\|E_{11}\|}{\|E\|}$	$\frac{\|E_{22}\|}{\|E\|}$	$\frac{\|E_{12}\|}{\|E\|}$	$\frac{\|E_{21}\|}{\|E\|}$	$\frac{\|V_1\|}{\|V\|}$	$\frac{\|V_2\|}{\|V\|}$
Syn. 1	10^5	99999	18.74	0.961	0.015	0.861	0.050	0.864	0.002	0.085	0.052	0.948
Syn. 2	10^5	99999	7.155	0.988	0.068	0.541	0.108	0.760	0.015	0.117	0.123	0.877
Syn. 3	25,000	46907	3.722	1.078	0.042	0.488	0.057	0.828	0.009	0.106	0.124	0.876

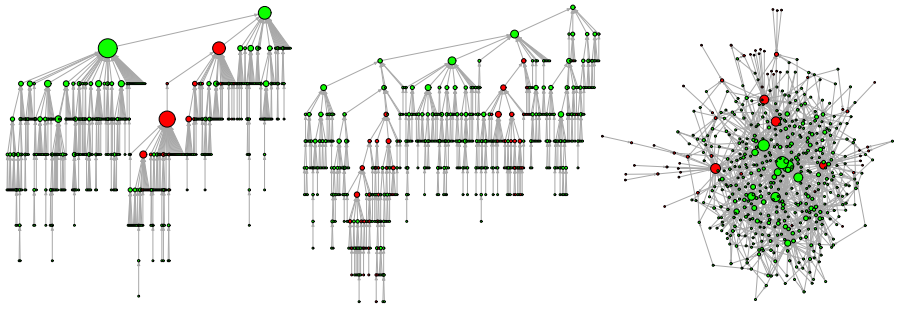

Fig. 1. Synthetic networks with 500 nodes: (l.h.s.) linear, tree network ($a = 0.003$, $D = 1$), (m.h.s.) sub-linear, tree network ($\alpha = 0.25$, $a = 0.02$, $D = 1$), (r.h.s.) linear, non-tree network ($m_1 = 1, m_2 = 2$, $a = 0.02$, $D = 1$). The red (green) circles represent the minority (majority) nodes with sizes proportional to the degrees.

Fig. 1 (l.h.s.). In the linear case, there are minority nodes with a large degree. For each sampling method, we estimate the probability of sampling a minority node through the proportion of minority nodes sampled over 10^4 runs. Additionally, we also compute the average of the degree-ranks and Page-ranks of the minority sampled nodes. The results are given in Table 2. Rank is expressed as percent in Table 2, where higher rank corresponds to smaller top percent. The probability of sampling a minority node under uniform sampling is close to the asymptotic value $\sqrt{a} \approx 0.055$ and does not give preference to "more popular" nodes (with higher degree or Page-rank). Sampling proportional to degree approximately doubles the chance to pick a minority node approaching $2\sqrt{a} - \frac{9}{2}a \approx 0.096$ and leads to a higher rank. The results improve with sampling proportional to in-degree which agrees with the asymptotic analysis. For sampling proportion to Page-rank (PR_c) with $c = m/(m+1)$, $m \in \mathbb{N}$, the mean number of walk steps is m. The number of steps being random does not improve the results. If the value of c is close to 0, PR_c is akin to uniform sampling. On the other hand, when c is large, the walk can hit the root. This can be explained by the diameter of the network which is 18 (in the tree case, it is $O(\log |V|)$). These drawbacks explain partly the good performance of fixed length walk sampling which also has the higher rank of the minority sampled nodes. This sampling scheme gives preference to nodes with a higher Page-rank as well.

We next consider the sub-linear, tree network with $\alpha = 0.25$ and $a = 0.02$ ($D = 1$) which gives $\pi_1 \approx 0.124$. The characteristics of the generated network are given in Table 1 (Syn. 2). An illustration of a small network with these characteristics is shown in Fig.1 (m.h.s.). We estimate the probability of sampling a minority and its importance for each sampling scheme using 10^4 runs – see

Table 2. Linear, tree network (Syn. 1): minority nodes representation

Sampling	U	D	ID	$PR_{1/2}$	$PR_{2/3}$	$PR_{3/4}$	$PR_{4/5}$	FL_2	FL_3	FL_4
prob.	0.052	0.110	0.133	0.077	0.090	0.093	0.089	0.189	0.191	0.150
degree-rank(%)	46.147	6.883	3.628	23.702	16.220	12.199	10.805	1.032	0.330	0.155
Page-rank(%)	46.215	7.323	3.912	23.759	16.199	12.195	10.781	0.825	0.212	0.080

Table 3. Sub-linear, tree network (Syn. 2): minority nodes representation

Sampling	U	D	ID	$PR_{2/3}$	$PR_{3/4}$	$PR_{4/5}$	$PR_{5/6}$	FL_4	FL_5	FL_6
prob.	0.125	0.176	0.226	0.199	0.220	0.226	0.223	0.387	0.401	0.381
degree-rank(%)	43.345	17.736	9.848	17.515	13.053	10.737	9.586	1.059	0.529	0.395
Page-rank(%)	43.037	18.954	10.417	17.284	12.783	10.553	9.358	0.617	0.384	0.143

Table 4. Linear, non-tree network (Syn. 3): minority nodes representation

Sampling	U	D	ID	$PR_{1/2}$	$PR_{2/3}$	$PR_{3/4}$	FL_2	FL_3	FL_4
prob.	0.1212	0.164	0.207	0.142	0.146	0.139	0.234	0.211	0.158
degree-rank(%)	69.045	17.184	9.443	46.636	35.758	29.978	3.211	1.283	0.609
Page-rank(%)	49.437	11.626	5.846	33.362	25.660	21.028	1.536	0.527	0.233

Table 3. The qualitative comparison of the performance of the sampling schemes is the same as in the linear case. However, the number of steps for sampling proportional to Page-rank and fixed length walk sampling is larger. The diameter of the generated network is 25.

Finally, we consider a linear, non-tree network with $m_1 = 1$ and $m_2 = 2$ and $a = 0.02$ ($D = 1$). The number of nodes is 25,000 which resulted in a network diameter of 16. The network properties are shown in Table 2 (Syn. 3) – see also Fig. 1 (r.h.s.) for a network generated with a smaller number of nodes. As seen from the results (averaged over 10^4 runs) in Table 4, the probability of sampling a minority node with fixed length walk sampling decreases compared to the sub-linear case due to the non-tree network structure (however, it is still approximately the double compared to uniform sampling).

3.4 Real Network

We inspect a social real-world network with a weak asymmetric homophily scenario to assess the probability of sampling a minority node. Hate is a retweet network where nodes denote users, and edges represent retweets among them. Users in the dataset are classified as either "hateful" (attribute 1) or "normal" (attribute 2) depending on the sentiment of their tweets [7]. "Hateful" users represent the minority. We consider the largest connected component of the network and remove loops and multiple edges for a comparison with the synthetic networks. Table 5 shows the key characteristics of interest of the directed network (with diameter 24). The results (averaged over 10^4 runs) in Table 6 are in line with the synthetic model, where fixed length walk sampling shows the higher probability of sampling a minority node in addition to a higher rank compared to uniform sampling. The smaller differences are due to the characteristics of

Table 5. Empirical network: characteristics

	$\|V\|$	$\|E\|$	D_1	D_2	H_{12}	H_{21}	$\frac{\|E_{11}\|}{\|E\|}$	$\frac{\|E_{22}\|}{\|E\|}$	$\frac{\|E_{12}\|}{\|E\|}$	$\frac{\|E_{21}\|}{\|E\|}$	$\frac{\|V_1\|}{\|V\|}$	$\frac{\|V_2\|}{\|V\|}$
Hate	2700	9709	9.976	0.579	0.408	0.529	0.333	0.386	0.122	0.158	0.183	0.817

Table 6. Hate network: minority nodes representation

Sampling	U	D	ID	$PR_{1/2}$	$PR_{2/3}$	$PR_{3/4}$	FL_2	FL_3	FL_4
prob.	0.179	0.199	0.205	0.194	0.204	0.205	0.214	0.224	0.222
degree-rank(%)	25.349	12.662	18.073	23.837	22.465	20.737	18.377	17.883	18.326
Page-Rank(%)	31.150	26.0812	15.363	27.259	23.328	20.579	13.911	13.272	14.227

the network, where the proportions of edges from "normal" to "hateful" users is only slightly higher than in the opposite direction. This can also be seen from the homophily measures H_{21} and H_{12}.

4 Relative Ranking of Minorities under Sampling

The aim of this section is to quantify the *relative ranking* of nodes of type 1 compared to type 2 by observing the attribute type counts in a pre-specified fraction $\gamma \in (0, 1)$ of nodes selected under one of the following two sampling schemes:

A: select γ fraction of nodes with the highest degrees (strictly speaking, this sampling does not involve randomness at the sampling level);
B: sample without (or with, but not used in the scenarios below) replacement γ fraction of nodes with probability proportional to degrees.

If an attribute type predominates the other attribute type in a given sampling scheme, we call it the *higher ranked attribute* for that scheme. As in Section 3, we thus consider the proportion of minority nodes in samples, but now focus on degree-based sampling schemes A and B, dependence on γ (for small sample sizes), and also on network structural properties such as homophily and out-degrees.

4.1 Synthetic Networks

For the synthetic network (1), we explore the questions above in terms of its model parameters κ (the propensity matrix determining homophily) and $\boldsymbol{m} = (m_1, m_2)$ (the out-degree vector). We shall gain insight through the following results and the quantities involved. From the analysis of the linear model [1], we have: as $n, k \to \infty$,

$$\widehat{\eta}_a^{\boldsymbol{m}} := \frac{\sum_{v \in V: a(v)=a} \deg(v, n)}{2(n + n_0)} \to \eta_a^{\boldsymbol{m}}, \quad \mathbf{p}_n^{\boldsymbol{m},a}(k) \sim k^{-(1+2/\phi_a^{\boldsymbol{m}})}, \; a = 1, 2, \qquad (2)$$

where $\eta_a^{\boldsymbol{m}}$ represents the limit of the normalized sum $\widehat{\eta}_a^{\boldsymbol{m}}$ of degrees of attribute type a and $\mathbf{p}_n^{\boldsymbol{m},a}(k)$ represents the proportion of nodes of type a with degree k which follows a power law with exponent $\Phi_a^{\boldsymbol{m}} := 2/\phi_a^{\boldsymbol{m}}$ in the limit. The quantities $\eta_a^{\boldsymbol{m}}$ and $\phi_a^{\boldsymbol{m}}$ are related to the relative ranking of minorities under the

Table 7. Heterophilic synthetic networks: proportion of minority nodes.

γ	0.01	0.02	0.03	0.04	0.05	0.1	0.15	0.20	0.3	0.4	0.5
Scheme A: $m_1 = 1, m_2 = 1$	0.868	0.810	0.733	0.673	0.626	0.514	0.467	0.417	0.407	0.355	0.325
$m_1 = 5, m_2 = 1$	0.372	0.458	0.515	0.560	0.576	0.660	0.679	0.701	0.732	0.754	0.513
Scheme B: $m_1 = 1, m_2 = 1$	0.442	0.432	0.433	0.422	0.422	0.397	0.384	0.371	0.355	0.343	0.334
$m_1 = 5, m_2 = 1$	0.535	0.5424	0.546	0.545	0.550	0.550	0.550	0.544	0.533	0.513	0.488

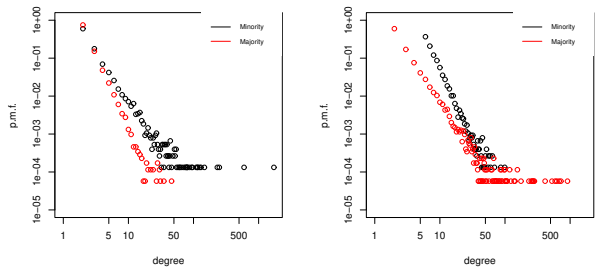

Fig. 2. Heterophilic networks (l.h.s.) $m_1 = m_2 = 1$ (r.h.s.) $m_1 = 5$, $m_2 = 1$: degree distribution.

two sampling schemes A and B above and can be precisely computed. $(\eta_1^{\boldsymbol{m}}, \eta_2^{\boldsymbol{m}})$ is the minimizer of a suitable function ([1], Eq. (4.1)) and

$$\phi_a^{\boldsymbol{m}} = 2 - m_a \pi_a / \eta_a^{\boldsymbol{m}}. \tag{3}$$

If $\phi_1^{\boldsymbol{m}} > \phi_2^{\boldsymbol{m}}$, the tail of the minority degree distribution is heavier (see Eq. (2)) and hence minorities are higher ranked in scheme A. On the other hand, if $\eta_1^{\boldsymbol{m}} > \eta_2^{\boldsymbol{m}}$, the probability of sampling a minority node is higher in each draw and hence the same conclusion holds in scheme B. We consider three different network configurations as follows (for the proofs of the results (4)-(9) below, see [1]).

Heterophilic Network. We first consider the scenario of a strongly heterophilic network, such that $\kappa_{11} = \kappa_{22} = 1$ and $\kappa_{12} = \kappa_{21} = K$ is large. In this case, node pairs with different attributes are more likely to be connected than node pairs with concordant attributes. As K increases, $\phi_1^{\boldsymbol{m}}$ and $\phi_2^{\boldsymbol{m}}$ behave as

$$\phi_1^{\boldsymbol{m}} \approx 2\left(1 - \frac{m_1\pi_1}{m_1\pi_1 + m_2\pi_2}\right), \qquad \phi_2^{\boldsymbol{m}} \approx 2\left(1 - \frac{m_2\pi_2}{m_1\pi_1 + m_2\pi_2}\right). \tag{4}$$

Thus, the rank of minority nodes under scheme A depends on the relation between $m_1\pi_1$ and $m_2\pi_2$. Table 7 shows the results for two linear networks with 25,000 nodes, $K = 10$ and $\pi_1 = 0.3$. The out-degree vectors $\boldsymbol{m}$ are $(1, 1)$ and $(5, 1)$. For $m_1 = 1$, we have $\phi_1^{\boldsymbol{m}} \approx 1.373$ and $\phi_2^{\boldsymbol{m}} \approx 0.659$ (using (3)) which are close, respectively, to 1.4 and 0.6 given by the approximations in (4). In this case $m_1\pi_1 < m_2\pi_2$, and the minority nodes rank higher under scheme A due to the fact that majority nodes tend to connect to minority nodes, increasing their ranks. This holds for any tree network. For $m_1 = 5$, we have $\phi_1^{\boldsymbol{m}} \approx 0.688$ and $\phi_2^{\boldsymbol{m}} \approx 1.377$ which are close, respectively, to 0.636 and 1.364 given by (4).

Table 8. Homogenous homophily networks ($m_1 = 5, m_2 = 1$; $m_1 = 5$, $m_1 = 2$) and homogenous mixing network ($m_1 = 2$, $m_2 = 1$): proportion of minority nodes.

γ	0.01	0.02	0.03	0.04	0.05	0.1	0.15	0.20	0.3	0.4	0.5
scheme A: $m_1 = 5, m_2 = 1$	0.888	0.886	0.887	0.881	0.890	0.868	0.859	0.857	0.86	0.751	0.601
$m_1 = 5$, $m_2 = 2$	0.700	0.718	0.697	0.693	0.694	0.692	0.698	0.658	0.673	0.6556	0.614
$m_1 = 2$, $m_2 = 1$	0.552	0.598	0.601	0.594	0.593	0.589	0.589	0.576	0.586	0.547	0.580
Scheme B: $m_1 = 5, m_2 = 1$	0.684	0.690	0.682	0.676 8	0.677	0.659	0.6441	0.629	0.596	0.558	0.516
$m_1 = 5$, $m_2 = 2$	0.527	0.522	0.518	0.522	0.516	0.505	0.497	0.493	0.473	0.455	0.436
$m_1 = 2$, $m_2 = 1$	0.5056	0.505	0.503	0.507	0.502	0.499	0.493	0.487	0.477	0.465	0.451

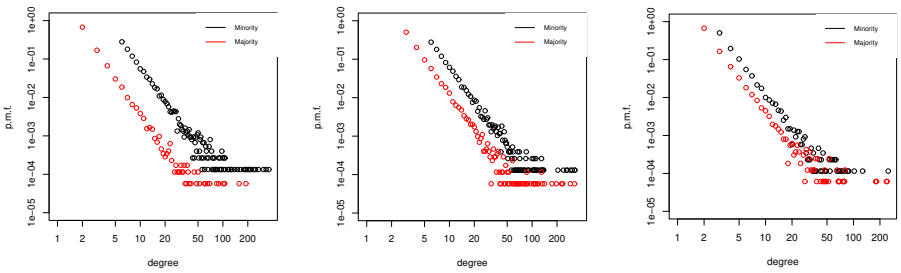

Fig. 3. Homogenous homophily (l.h.s.) $m_1 = 5$, $m_2 = 1$ (m.h.s.) $m_1 = 5$, $m_2 = 2$ and homogenous mixing: (r.h.s.) $m_1 = 2$, $m_2 = 1$: degree distribution.

In this setting $m_1\pi > m_2\pi_2$, the minority nodes increase the ranks of majority nodes for small values of γ, by connecting to the majority with more output edges. (Note that when $\gamma = 1$ the relative ranking is given by the proportion of minority nodes in the network.) Fig. 2 shows the degree distribution for each attribute, where in (l.h.s.) the minority has a heavier tail ($\phi_1^{\boldsymbol{m}} > \phi_2^{\boldsymbol{m}}$) and in (r.h.s.) it is the majority ($\phi_1^{\boldsymbol{m}} < \phi_2^{\boldsymbol{m}}$).

As K gets larger, $\eta_1^{\boldsymbol{m}}$ and $\eta_2^{\boldsymbol{m}}$ approach the same limit value

$$\eta_1^{\boldsymbol{m}} \approx \eta_2^{\boldsymbol{m}} \approx \frac{m_1\pi_1 + m_2\pi_2}{2}, \tag{5}$$

which implies that the differences between the relative rankings are smaller between the two attributes for scheme B. Table 7 shows the relative ranking of the minority for the networks described above under this scheme (the results were averaged over a large number of runs). For $m_1 = 1$, we have $\eta_1^{\boldsymbol{m}} \approx 0.478$ and $\eta_2^{\boldsymbol{m}} \approx 0.522$ which are close to 0.5 given by the approximation in (5). For $m_1 = 5$, we have $\eta_1^{\boldsymbol{m}} \approx 1.144$, $\eta_2^{\boldsymbol{m}} \approx 1.056$ which approach 1.1 in (5). However, the higher value of $\eta_1^{\boldsymbol{m}}$ makes the minority slightly more dominant for scheme B.

Homogenous Homophily and Homogenous Mixing. We consider the cases of a strong homogeneous homophily with $\kappa_{21} = \kappa_{21} = 1$ and $\kappa_{11} = \kappa_{22} = K$ large and homogenous mixing with all the elements of the matrix κ equal to 1. As K goes to infinity, the exponents of the tail degree distribution per attribute are equal and behave as

$$\phi_1^{\boldsymbol{m}} = \phi_2^{\boldsymbol{m}} \approx 1, \tag{6}$$

Table 9. Asymmetric homophily: proportion of minority nodes.

γ		0.01	0.02	0.03	0.04	0.05	0.1	0.15	0.20	0.3	0.4	0.5
Scheme A:	$m_1 = 1, m_2 = 1$	0.712	0.632	0.589	0.570	0.526	0.468	0.4565	0.397	0.384	0.345	0.323
	$m_1 = 2, m_2 = 1$	0.980	0.938	0.911	0.890	0.866	0.769	0.725	0.702	0.597	0.601	0.596
Scheme B:	$m_1 = 1, m_2 = 1$	0.423	0.411	0.396	0.392	0.390	0.370	0.363	0.357	0.342	0.334	0.327
	$m_1 = 2, m_2 = 1$	0.591	0.585	0.568	0.567	0.559	0.539	0.520	0.501	0.4766	0.449	0.425

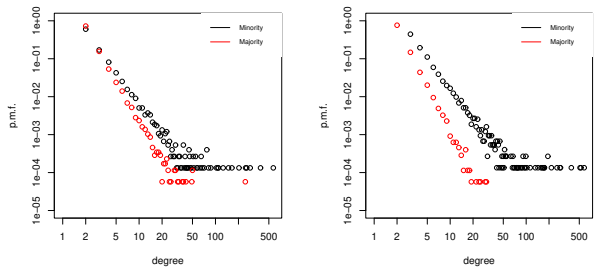

Fig. 4. Asymmetric Homophily: (l.h.s.) $m_1 = m_2 = 1$ (r.h.s.) $m_1 = 2$, $m_2 = 1$: degree distribution.

which also holds in the case of homogenous mixing. However, we will see that the relative ranking of the minority under scheme A will depend on the ratio m_1/m_2. Table 8 depicts two homogenous homophily networks with 25,000 nodes, $K = 10$, $\pi_1 = 0.3$, and $\boldsymbol{m}$ vectors $(5, 1)$ and $(5, 2)$ which result in $\phi_1^{\boldsymbol{m}} \approx 1.022$, $\phi_2^{\boldsymbol{m}} \approx 0.948$ and $\phi_1^{\boldsymbol{m}} \approx 1.003$, $\phi_2^{\boldsymbol{m}} \approx 0.997$, respectively. An homogenous mixing network with 25,000 nodes, $\pi_1 = 0.35$ and $\boldsymbol{m} = (2, 1)$ is also considered. Fig. 3 shows the degree distributions per attribute. Despite the degree tail exponents being similar from the plots, if m_1 is larger than m_2, the degrees of minority nodes get a high initial boost. Additionally, from the works [4, 8], for multi-attributes, there is a "persistence phenomenon", i.e., the maximal degree nodes from any attribute type emerge from, with high probability, the oldest nodes of that type added to the network. Therefore, the results in Table 8 show that minority nodes have a higher ranking under scheme A.

On the other hand, as K goes to infinity (homogeneous homophily) and also for homogenous mixing,

$$\eta_1^{\boldsymbol{m}} \approx m_1\pi_1, \qquad \eta_2^{\boldsymbol{m}} \approx m_2\pi_2. \tag{7}$$

For the networks considered with $\boldsymbol{m} = (5, 1)$ and $\boldsymbol{m} = (5, 2)$, the exact values (resp. approximations in (7)) are $\eta_1^{\boldsymbol{m}} \approx 1.534$, $\eta_2^{\boldsymbol{m}} \approx 0.666$ (resp. 1.5 and 0.7), and $\eta_1^{\boldsymbol{m}} \approx 1.504$, $\eta_2^{\boldsymbol{m}} \approx 1.396$ (resp. 1.5 and 1.4). For $\boldsymbol{m} = (2, 1)$, the true value and approximation match with $\eta_1^{\boldsymbol{m}} \approx 0.7$, $\eta_2^{\boldsymbol{m}} \approx 0.65$. Thus, if $m_1\pi_1 > m_2\pi_2$, the minority nodes rank higher under scheme B – see Table 8.

In both types of networks, minority nodes can increase their popularity via schemes A and B through a higher ratio m_1/m_2. In the context of social networks, it means minorities increasing their social interaction.

Asymmetric Homophily. The last scenario is the case of a strong asymmetric homophily network (slightly different from Sec. 3), where $\kappa_{11} = K$ is large, and

Table 10. Empirical networks: proportion of minority nodes.

γ		0.01	0.02	0.03	0.04	0.05	0.1	0.15	0.20	0.3	0.4	0.5
Scheme A:	Hate	0.778	0.778	0.716	0.704	0.689	0.637	0.588	0.533	0.430	0.356	0.304
	APS	0.154	0.115	0.132	0.157	0.203	0.211	0.250	0.266	0.297	0.289	0.300
Scheme B:	Hate	0.472	0.460	0.460	0.460	0.451	0.427	0.413	0.404	0.386	0.311	0.282
	APS	0.269	0.294	0.277	0.278	0.302	0.291	0.284	0.283	0.300	0.298	0.305

$\kappa_{22} = \kappa_{12} = \kappa_{21} = 1$. As K tends to infinity,

$$\phi_1^{\boldsymbol{m}} \approx \frac{2m_1\pi_1 + 3m_2\pi_2}{2m_1\pi_1 + 2m_2\pi_2}, \quad \phi_2^{\boldsymbol{m}} \approx \frac{m_2\pi_2}{m_1\pi_1 + m_2\pi_2} \tag{8}$$

and

$$\eta_1^{\boldsymbol{m}} \approx \frac{2m_1\pi_1(m_1\pi_1 + m_2\pi_2)}{2m_1\pi_1 + m_2\pi_2}, \quad \eta_2^{\boldsymbol{m}} \approx \frac{m_2\pi_2(m_1\pi_1 + m_2\pi_2)}{2m_1\pi_1 + m_2\pi_2}. \tag{9}$$

Two networks are considered with 25,000 nodes, $K = 10$ and $\boldsymbol{m}$ vectors $(1, 1)$ and $(2, 1)$ in Table 9. In both networks, $\phi_1^{\boldsymbol{m}} > \phi_2^{\boldsymbol{m}}$ and the minorities rank higher under scheme A. The exact values are $\phi_1^{\boldsymbol{m}} \approx 1.31$, $\phi_2^{\boldsymbol{m}} \approx 0.761$ ($m_1 = 1$) and $\phi_1^{\boldsymbol{m}} \approx 1.247$, $\phi_2^{\boldsymbol{m}} \approx 0.609$ ($m_1 = 2$) which are close to the approximations in (8). This also agrees with the degree tail exponents in Fig. 4 with the degree distribution of the minority being more heavy-tailed (higher $\phi_1^{\boldsymbol{m}}$).

Under scheme B, minorities rank higher with $m_1 = 2$ since $2m_1\pi_1 > m_2\pi_2$ in (9) (also $\eta_1^{\boldsymbol{m}} \approx 0.821$, $\eta_2^{\boldsymbol{m}} \approx 0.479$). This means that in a social network, if the arriving majority nodes have almost a neutral attribute preference attachment ($\kappa_{12} = \kappa_{22}$), the minorities can increase their popularity through the number of outgoing edges that connect to other minority nodes.

4.2 Real Networks

We consider two real-world networks with power-law degree distributions to assess the ranking of the minorities under schemes A and B. For the Hate network in Section 3.4, the exponents of the fitted degree distributions ($\widehat{\Phi}_a^{\boldsymbol{m}}$) are 2.776 and 3.338; and the normalized sums of the degrees ($\widehat{\eta}_a^{\boldsymbol{m}}$) are 3.223 and 3.617 for the minority and majority, respectively. Table 10 shows that under scheme A, the minorities rank higher. APS is a scientific network from the American Physical Society where nodes represent articles from two subfields and edges represent citations with homogeneous homophily. Some networks statistics are: 1281 (nodes), 3064 (edges). The minority rank is lower in both schemes where the exponents and normalized sums of the degrees are 3.947 and 3.292, and 1.332 and 3.452 respectively, for subfields 1 (minority) and 2 (majority). For these two real networks, the results on relative ranking of the minority are in line with those for the synthetic networks.

5 Conclusions and Future Work

This paper explored settings where Page-rank and walk-based network sampling schemes favor small minority attribute nodes compared to uniform sampling. We also investigated the conditions for the minority nodes to rank higher in degree-based sampling. To this end, we used an attributed network model with

homophily under several network configurations which provided insight into real-world networks.

In follow-up work, we plan to compare and contrast the performance of various centrality measures, including degree and Page-rank centrality, for ranking and attribute reconstruction tasks in the semi-supervised setting, where one has partial information on the attributes and wants to reconstruct it for the rest of the network. In the setting of dynamic and evolving networks, contrary to static networks, preliminary results in [1] seem to suggest starkly different behavior between degree vs Page-rank centrality in such settings.

Acknowledgements. S.Ba. is partially supported by the NSF CAREER award DMS-2141621. S.Bh. and V.P. are partially supported by NSF DMS-2113662. S. Ba., S.Bh. and V.P. are partially supported by NSF RTG grant DMS-2134107.

References

1. N. Antunes, S. Banerjee, S. Bhamidi, and V. Pipiras. Attribute network models, stochastic approximation, and network sampling and ranking. *Preprint arXiv:2304.08565v1*, 2023.
2. 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.
3. 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.
4. S. Banerjee and S. Bhamidi. Persistence of hubs in growing random networks. *Probability Theory and Related Fields*, 180(3-4):891–953, 2021.
5. P. Chebolu and P. Melsted. Pagerank and the random surfer model. In *SODA*, volume 8, pages 1010–1018, 2008.
6. F. W. Crawford, P. M. Aronow, L. Zeng, and J. Li. Identification of homophily and preferential recruitment in respondent-driven sampling. *American Journal of Epidemiology*, 187(1):153–160, 2018.
7. L. Espín-Noboa, C. Wagner, M. Strohmaier, and F. Karimi. Inequality and inequity in network-based ranking and recommendation algorithms. *Scientific Reports*, 12(1):1–14, 2022.
8. P. Galashin. Existence of a persistent hub in the convex preferential attachment model. *arXiv preprint arXiv:1310.7513*, 2013.
9. F. Karimi, M. Génois, C. Wagner, P. Singer, and M. Strohmaier. Homophily influences ranking of minorities in social networks. *Scientific Reports*, 8(1):1–12, 2018.
10. M. G. Merli, A. Verdery, T. Mouw, and J. Li. Sampling migrants from their social networks: The demography and social organization of Chinese migrants in Dar es Salaam, Tanzania. *Migration Studies*, 4(2):182–214, 2016.
11. T. Mouw and A. M. Verdery. Network sampling with memory: a proposal for more efficient sampling from social networks. *Sociological Methodology*, 42(1):206–256, 2012.
12. 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.
13. A. Stolte, G. A. Nagy, C. Zhan, T. Mouw, and M. G. Merli. The impact of two types of COVID-19-related discrimination and contemporaneous stressors on Chinese immigrants in the US South. *SSM-Mental Health*, 2:100159, 2022.