Social Network Analysis and Mining (2024) 14:158
https://doi.org/10.1007/s13278-024-01326-6

ORIGINAL ARTICLE

Representation, ranking and bias of minorities in sampling attributed networks

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

Received: 30 April 2024 / Revised: 12 July 2024 / Accepted: 29 July 2024

Abstract
We investigate three related problems concerning sampling minorities in attributed networks. This is guided by a general attributed network model which can incorporate several levels of homophily and heterophily, and whose degree and Page-rank distributions have known properties. The first problem investigates sampling schemes that favor the representation of the minority over majority nodes and give preference to "more popular" minority nodes (i.e. higher degree/Page-rank) for a given homophily scenario. We show that (in-)degree and Page-rank sampling schemes increase the probability of sampling a minority node. The second problem concerns the relative ranking of minorities compared to majorities in degree and Page-rank based sampling schemes for several homophily and heterophily scenarios. We provide analytical conditions for the minority nodes to rank higher as a function of the model parameters for the degree based samplings and investigate the problem numerically for Page-rank based sampling schemes. The third problem considers subgraph sampling schemes and the bias of the proportion of minority nodes in top ranked degree nodes in several homophily and heterophily scenarios. Finally, the results and findings obtained from the sampling analysis are assessed on real-world networks.

Keywords Random networks · Attributes · Homophily · Heterophily · Sampling · Minorities · Ranking

1 Introduction

Attributed networks are graphs in which nodes (or edges) have attributes (features). In real-world networks, the attributes across connections will co-vary and are not independent. One standard phenomenon in many such real world systems is homophily (Shrum et al. 1988; McPherson et al. 2001; Mislove et al. 2010), i.e., node pairs with similar attributes being more likely 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. A contrasting co-variation phenomenon is heterophily, where nodes with similar attributes (or the same type) "repel" each other. 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). The networks are further used for ranking individuals according to their centrality scores (measured via functionals such as degree or Page-rank scores) which further exacerbate inequalities in representation of minorities in the network through algorithms such as recommendation systems that use the underlying network structure (Espín-Noboa et al. 2022), or effect the flow of information and the perceptions of minorities within the network (Lee et al. 2019). Another major direction for understanding the role of attributes is the maximization of influence problem. Since the pioneering work (Granovetter 1978), followed by the

Sayan Banerjee, Shankar Bhamidi, Vladas Pipiras contributed equally to this work.

✉ Nelson Antunes
nantunes@ualg.pt

Sayan Banerjee
sayan@email.unc.edu

Shankar Bhamidi
bhamidi@email.unc.edu

Vladas Pipiras
pipiras@email.unc.edu

[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

Published online: 10 August 2024

path-breaking (Kempe et al. 2003) in the setting of computer science and combinatorial optimization, the main goal has been to understand, in the context of viral marketing, which set of individuals in the network to seed, with information or a product, so as to maximize its spread. While Kempe et al. (2003) has lead to a thriving research direction, in the last two years there has been significant realization as well as understanding that ignoring attributes of individuals within the network, and in particular the impact of homophilly in the connectivity as well as strength of ties between individuals, can significantly hamper the efficacy of proposed algorithms as well as conclusions (Aral and Walker 2012; Aral and Dhillon 2018) both in the context of empirical systems (Caliò and Tagarelli 2021) and even in the context of the standard benchmarking pipelines used to check the performance of algorithms (Sziklai and Lengyel 2022, 2024). Further Caliò and Tagarelli (2021) tackles the significant challenges around the attribute diversity of the seed set for influence maximization via approximation schemes for submodular functions. Optimizing such functions in practice leads to running diffusion schemes via sampling from the underlying node set according to specified distributions and then running influence cascades from these nodes.

Attributed network models play a major role in understanding the impact of the network evolution with homophily/heterophily and preferential attachment in the representation of the minority group. Here, *homophily* (Shrum et al. 1988; McPherson et al. 2001; Mislove et al. 2010) corresponds to the fundamental finding in many social network settings of node pairs with similar attributes being likelier connected than node pairs with discordant attributes. A precise quantitative version of this measure is given in Sect. 2.1. *Preferential attachment* (Barabási and Albert 1999) refers to the notion that when new nodes enter a networked system, they tend to connect to pre-existing nodes with probability proportional to some monotonically increasing function of the degree of the existing nodes, thus reinforcing the popularity of current nodes. Once again, a precise definition is given in Sect. 2.1. In this context, the use of models were initiated in Karimi et al. (2018), where the authors used fluid limit analysis to study the limiting degree distributions for two attributes (minority and majority). Through numerical simulations, they showed the effect of homophily and heterophily in reducing or amplifying the ranking of minority nodes in the network according to their degree. Inequality for Page-rank scores centrality measure and the representation of minority amongst high ranking nodes were studied in Espín-Noboa et al. (2022) using a similar model in the case of a directed network where nodes can became active to connect to other nodes. In both works, the impact of sampling in the ranking of minority nodes was not considered.

Given that large networks can only be partially observed, sampling has been an activate area of research across different subjects (see e.g. Antunes et al. 2021a, b and the references therein). Initial research on sampling has shown that conclusions from samples depend on network properties (e.g. scale free), the characteristic of the measure of interest (e.g. degree), and the sampling method and rate used (Leskovec and Faloutsos 2006). A related question is whether sampling preserves the representation/ranking of minority nodes, or perhaps increases their visibility in the sample, when compared with the whole network. Sampling in networks with homophily/heterophily has received little attention in the literature. The bias of classical sampling methods in preserving the ranking of nodes and visibility of minorities under a similar model as in Karimi et al. (2018) was investigated in Wagner et al. (2017). However, the analysis was based only on empirical results. In a different direction (Espín-Noboa et al. 2021), synthetic models are used to understand the accuracy of prediction of attribute labels given partial information of the labels of a subset of seeded nodes; the goal is to understand the impact of homophily/heterophily and preferential attachment driven growth characteristics of the underlying network on the accuracy of classifiers and inference algorithms. In Antunes et al. (2023b), random walk sampling algorithms are considered to infer several functionals of attribute networks such as homophily/heterophily measures, attribute and degree distributions per attribute.

The aim of this paper is to provide analytical and numerical results for three related problems concerning representation, ranking and bias of minorities based on the degree and Page-rank centrality measures in sampling attributed networks (extending the knowledge in the literature Karimi et al. 2018; Espín-Noboa et al. 2022; Wagner et al. 2017).

To this end, we consider a dynamic random directed network model generalizing (Karimi et al. 2018) where each arriving node connects to a fixed number of nodes (outgoing edges) depending on its attribute. The probability that each edge connects to a node of the network is proportional to its degrees (raised to the power of a parameter $\alpha \geq 0$) and a function that measures the propensity of the attributes of the nodes to interact. This allows to represent the two main mechanisms of the formation found in social networks: preferential attachment ($\alpha > 0$) and homophily/heterophily. We give analytical results for the degree and Page-rank distributions per attribute in the setting where popularity depends in a linear fashion on the current number of connections of a node (the regime $\alpha = 1$) as the size of the network increases. In general, the models considered in this paper, with self-reinforcement, where nodes with high degree have a higher propensity to obtain future connections, are non-trivial to analyze analytically; in the specific regime $\alpha = 1$, it turns out (Antunes et al. 2023a; Jordan 2013) that network functionals can be derived for the degree and Page-rank distributions.

The general sublinear case, where popularity of nodes is a sublinear function of the current degree ($\alpha \in (0, 1)$) in the context of attributed network models, is open till date, and is studied numerically in this paper. The results imply that while degree distribution tail exponents depend on the attribute type, Page-rank score distributions have the same tail exponent across attributes; thus in the context of extremal behavior or most "popular nodes", measuring the centrality of nodes using their degree is much more affected by attribute information, than a more global Page-rank centrality; for example in the hompophilic regime, while minority attributes are automatically disadvantaged by degree centrality, this is not the case with Page-rank centrality. Moreover, the mean behavior of the limiting Page-rank score distributions can be explicitly described and shown to depend on the attribute type. We use the model in the case of minority and majority nodes to investigate:

(a) *Sampling a Rare Minority*: An area of significant research interest in the context of network sampling comprises settings where there is a particular rare minority which has higher propensity to connect within itself as opposed to majority nodes; for substantial recent applications and impact of such questions, see (Mouw and Verdery 2012; Merli et al. 2016; Stolte et al. 2022). 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. Sampling schemes such as uniform sampling often struggle to find a non-trivial proportion of minority nodes for further downstream sociological explorations; see for example (Stolte et al. 2022; Merli et al. 2016) for questions related to mental health questions and demographic or social profiles, related to rare minorities within large populations. 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. This leads us to analyze several sampling schemes based on (in-) degree and Page-rank centrality measures. We quantify explicitly the probability of sampling a minority node in a *linear network* ($\alpha = 1$) in the case that each arrival node connects to only one node of the network (i.e. *tree network*) and investigate the problem numerically for other network configurations (non-linear and non-tree networks). The results show that sampling schemes based on Page-rank centrality increase the probability of sampling a minority node and its "popularity" (higher degree and Page-rank).

(b) *Centrality-Based Sampling and Higher Ranking of Minorities*: We consider sampling schemes based on the degree and Page-rank centrality and investigate conditions for the minority nodes to rank higher (i.e. the proportion of the minority nodes in the sample is higher than for the majority nodes). For the degree centrality we provide explicit conditions for higher rank when a small fraction of nodes is sampled as a function of the model parameters (node attribute probabilities and out-degrees) in a linear network and several network scenarios: heterophily, homogeneous homophily (homogenous mixing) and asymmetric homophily. For the Page-rank centrality the results are investigated numerically and provide insights for the minority nodes to rank higher in the same scenarios.

(c) *Bias of Subgraph Samplings in Ranking Through Degree Centrality*: We consider a different sampling schemes from (b) where nodes (resp., edges) are sampled and the induced (resp., incident) subgraph is observed. The goal is to measure the bias of sampled subgraph in the proportion of the minority nodes in the top percentile of high degree nodes. For the tree linear network we provide an analytical result to compute the bias for induced subgraph sampling. The sign of the bias which represents under or over representation of minorities in the subgraph samplings is then investigated numerically for the homophily and heterophily scenarios as a function of the model parameters.

The details of the derivation of the analytical results for the special network configurations using stochastic approximations are defered to the technical report (Antunes et al. 2023a). Finally, the analytical and numerical results are assessed on real-world networks with several levels of homophily and heterophily showing a good agreement with the findings of the sampling analyses in the considered network model.

This paper is a significant extension of the conference paper (Antunes et al. 2024) including: (1) a new Sect. 3 with the limiting distributions of the degree and Page-rank measures per attribute and their properties which are numerically illustrated for finite size networks; (2) the results of the degree centrality based sampling and higher ranking of minorities in Sect. 5.1 have been extended to several network configurations which are now visualized through plots; (3) a new Sect. 5.2 is included with Page-rank centrality based sampling and higher ranking of minorities; (4) a new Sect. 6 is added that investigates the bias of subgraph samplings in ranking through degree centrality measures; (5) a separate section with real-world networks (Sect. 7) including additional datasets provides evidence of the similarities to the considered model and the network sampling analyse; (6) finally, parts of the remaining sections have been improved

including an algorithm to generate the attributed dynamical model in Sect. 2.1.

2 Preliminaries

In this section, we introduce the network model, homophily and heterophily measures, and the network scenarios considered in our experiments.

Algorithm 1 Dynamic Attributed Network Model

Input: connected network G_0 with node labels $1, \ldots, n_0$ and attributes $a(1), \ldots, a(n_0)$
Initialization: $\kappa(=(\kappa_{a,b})_{a,b\in A})$; $(m_a)_{a\in A}$; α; n (number of arrival nodes)
for $i = 1$ **to** n **do**
  $a(n_0+i) \leftarrow$ select an attribute in A with weights $(\pi_a)_{a\in A}$
  for $j = 1$ **to** $m_{a(n_0+i)}$ **do**
    $v[j] \leftarrow$ select a node from G_{i-1} with probabilities proportional to
      $\kappa_{a(v),a(n_0+i)}[\deg(v)]^\alpha, v \in G_{i-1}$
  end for
  $G_i \leftarrow$ add node n_0+i to G_{i-1}
  for $j = 1$ **to** $m_{a(n_0+i)}$ **do**
    $G_i \leftarrow$ add edge $(n_0+i, v[j])$ to G_{i-1}
  end for
end for
Output: G_n and $(a(n_0+i))_{1\le i\le n}$

2.1 Dynamic attributed network model

We describe a network model where nodes have attributes which modulate the evolution dynamics of the network. This will have impact on the network structure and thus on the ranking of the nodes based on the network centrality measures. Let $A = \{1, 2, \ldots, L\}$ be a finite set of the attribute labels. We describe the dynamics of a sequence of growing networks from an initial state. At time 0, a base connected directed network G_0 with n_0 nodes is given, where every node $v \in G_0$ has an attribute $a(v)$ in A. At each discrete time $n = 1, 2, \ldots$, a node enters the network. The probability that an arriving node has attribute a is π_a independent of the current network. An entering node of attribute type a connects to the network through $m_a \geq 1$ outgoing edges. The propensity with which a node with attribute b attaches to a node with attribute a is given by $\kappa_{a,b}$. Let $\kappa(=(\kappa_{a,b})_{a,b\in A})$ be the propensity matrix. Additionally, let $\alpha \geq 0$ be the preferential attachment parameter associated with the strength of popularity of a node. A node u that arrives at time n, connects any of its $m_{a(u)}$ edges independently to a node $v \in G_{n-1}$ according to

$$\mathbb{P}(u \to v | G_{n-1}, a(u)) \propto \kappa_{a(v),a(u)}[\deg(v)]^\alpha, \tag{1}$$

where $\deg(v)$ is the degree of node v at time $n-1$ (if $|G_0| = 1$, then we set $\deg(v) = 1$). The attachment probabilities capture the combined effect of attribute types and node popularity in network evolution. A description of the algorithm to construct the dynamic network is given in Algorithm 1.

The model includes several classes of network dynamics. For instance, if $\kappa_{a,b} = 1$ for all $a, b \in A$, then there is no dependence of the attributes on the evolution of the network. In this case, if $\alpha = 1$, we have the classical Barabási-Albert model (Barabási and Albert 1999) (*linear* with $\alpha = 1$) and with $0 < \alpha < 1$ the *sublinear* preferential attachment model. When $\alpha = 0$ the incoming nodes attach to pre-existing nodes based purely on their attribute and are agnostic to the degree information—*uniform attachment model*.

2.2 Dyadicity and heterophilicity measures

The proposed model incorporates several features found in real world social networks such as homophily and heterophily. There are several ways to measure these characteristics of networks. Here, we use dyadicity and heterophilicity proposed in Park and Barabási (2007) for signed networks and that conveniently apply to directed networks. For a directed network, let V and E denote, respectively, the set of nodes and edges of the network. Let also V_a represent the set of nodes with attribute a and E_{ab} the set of directed edges from

nodes with attribute a to nodes with attribute b. Dyadicity for an attribute is defined as

$$D_a = |E_{aa}|/(|V_a|(|V_a| - 1)p), \quad a \in A,$$

where $p = |E|/(|V|(|V| - 1))$ is the edge density. Heterophilicity H_{ab} is given by

$$H_{ab} = |E_{ab}|/(|V_a||V_b|p), \quad a, b \in A, a \neq b.$$

Dyadicity and heterophilicity measure, respectively, the connectedness between nodes with the same and different attributes when compared to a random configuration of the network (i.e., when all edges are randomly distributed). If $D_a > 1$ and $H_{ab} < 1$, then nodes with attribute a attract each other and connections from nodes with attribute a to nodes with attribute b are repelled (homophily). On the other, $D_a < 1$ and $H_{ab} > 1$ represent heterophily. These quantities can be asymmetric among node attributes. Illustrations of the network structures generated with Algorithm 1 for linear and uniform attachment model for two attributes with homophily and heterophily are shown in Fig. 1. Throughout the paper, nodes with *attribute* 1 will be referred as *minority* and *attribute* 2 as *majority*. The synthetic datasets are generated and experiments are conducted in this paper with R package igraph (Csardi and Nepusz 2006). A summary of the main notation is given in Table 1.

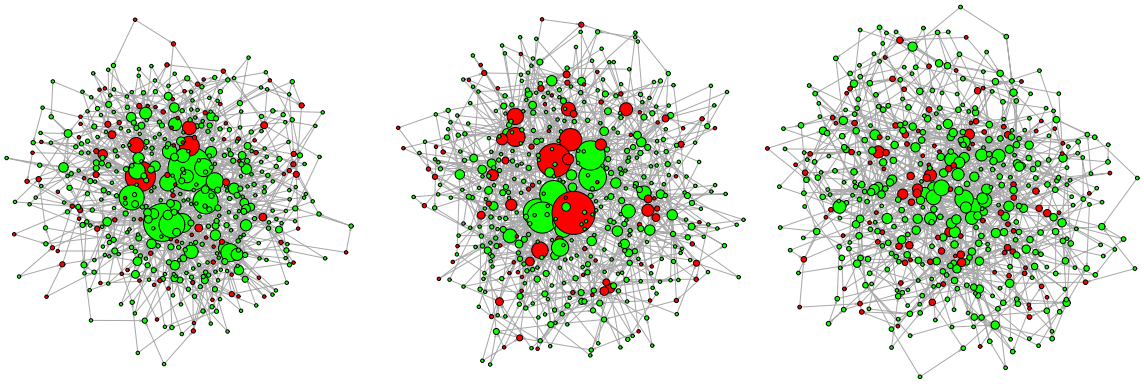


Fig. 1 Networks generated with the dynamic attributed network model (500 nodes, $\pi_1 = 0.2$): (left) $\alpha = 1$, homophily $\kappa_{11} = \kappa_{22} = 2$, $\kappa_{12} = \kappa_{21} = 1$; (middle) $\alpha = 1$, heterophily $\kappa_{11} = \kappa_{22} = 1$, $\kappa_{12} = \kappa_{21} = 2$; (right) $\alpha = 0$, homophily $\kappa_{11} = \kappa_{22} = 2$, $\kappa_{12} = \kappa_{21} = 1$. The red circles represent attribute 1 (minority) nodes and green attribute 2 (majority) nodes with sizes proportional to the degrees. The dyadicity and heterophilicity measures are: (left) $D_1 = 1.652$, $D_2 = 1.109$, $H_{12} = 0.838$, $H_{21} = 0.56$; (middle) $D_1 = 0.846$, $D_2 = 0.699$, $H_{12} = 1.038$, $H_{21} = 2.190$; (right) $D_1 = 1.613$, $D_2 = 1.109$, $H_{12} = 0.847$, $H_{21} = 0.564$ (color figure online)

Table 1 Summary of the main notation

Notation	Description
G_n	Graph at time n generated with the dynamic attributed network model
A	Set of attribute labels
π_a	Probability of an arriving node having attribute a
m_a	Number of edges a node with attribute a entering the network connects to pre-existing nodes
$\kappa_{a,b}$	Propensity of node with attribute b to connect to node with attribute a
α	Preferential attachment parameter
$\deg(v)$	Degree of node v
V	Set of nodes of the network
E	Set of edges of the network
V_a	Set of nodes of the network with attribute a
E_{ab}	Set of directed edges from nodes with attribute a to nodes with attribute b
D_a	Dyadicity of nodes with attribute a
H_{ab}	Heterophilicity from nodes with attribute a to nodes with attribute b
$p_a(k)$	(limit) Probability that a node with attribute a has degree k
$R_c(v)$	Page-rank score of node $v \in G_n$ with damping factor c and $\overline{R}_c(v) := nR_c(v)$
$\|B\|$	Number of elements of set B

2.3 Network homophily and heterophily scenarios

In our experiments, we consider mainly three network scenarios using different configurations of the propensity matrix κ. The *heterophily scenario* assumes $\kappa_{11} = \kappa_{22} = 1$ and $\kappa_{12} = \kappa_{21} = K$, where K is large. The *homogenous homophily scenario* takes $\kappa_{12} = \kappa_{21} = 1$ and $\kappa_{11} = \kappa_{22} = K$ for large K. The last configuration is the *asymmetric homophily scenario*, namely, $\kappa_{11} = K \gg 1$ and $\kappa_{22} = \kappa_{12} = \kappa_{21} = 1$. We let the proportion of minorities π_1 be small and investigate tree networks ($m_1 = m_2 = 1$) and non-tree networks ($m_1 > 1$ or $m_2 > 1$).

3 Distributions of centrality measures

In this section, we discuss analytical and numerical properties of the asymptotic distributions of the degree and Pagerank per attribute of the network model. These two centrality measures are used to sample and rank nodes in the following sections. The details (proofs) of the theoretical results can be found in Antunes et al. (2023a).

3.1 Degree distribution

For a linear network G_n ($\alpha = 1$), as n tends to infinity, the limiting probability mass function (p.m.f.) of the degree of nodes with attribute a is given by

$$p_a(k) = \frac{2}{\phi_a} \frac{\Gamma\left(m_a + \frac{2}{\phi_a}\right)\Gamma(k)}{\Gamma\left(k + 1 + \frac{2}{\phi_a}\right)\Gamma(m_a)}, \qquad k \geq m_a,$$

where Γ denotes the gamma function and

$$\phi_a = 2 - \frac{m_a \pi_a}{\eta_a}. \tag{2}$$

The quantity ϕ_a carries no special meaning and enters into the tail exponent in (3) below. The quantity η_a can be interpreted as the normalized sum of the degrees of nodes with attribute a as n tends to infinity and can be computed explicitly [η_a, $a \in A$, are the minimizers of a function given in Antunes et al. (2023a), Equation (4.1)]. For each $a \in A$, we have

$$p_a(k) \sim k^{-(1+2/\phi_a)}, \text{ as } k \text{ tends to infinity}. \tag{3}$$

The result implies that the limiting degree distribution follows a power-law with exponent $2/\phi_a$ dependent on the attribute. This agrees with the empirical evidence also found in real-world social networks (see Sect. 7).

In contrast, in the case of the uniform attachment model ($\alpha = 0$) and $m_a = 1$ (for simplicity, although the result can be extended to $m_a \geq 2$), the limiting degree distribution of attribute a is geometric with parameter $1/(1 + \phi_a)$ and has exponential tail.

Figure 2 shows the empirical degree distribution of G_n and the limiting distribution for several attributed networks, with parameters specified in the figure caption. We consider linear and uniform attachment networks with homophily and attributes 1 (minority) and 2 (majority). In all the cases, the bulk of the distribution per attribute type is approximated well by the limiting distribution. (We note that horizontal points for large degree values in the empirical distributions are due to the effect of the finite size of the network.) For the linear network ($\alpha = 1$) with $m_1 = m_2 = 1$ (Fig. 2, left), the maximum likelihood estimates of the empirical tail exponents are 2.347 and 1.734 which are close, respectively, to $2/\phi_1 \approx 2.566$ and $2/\phi_2 \approx 1.917$ [given by (2)], where the majority attribute has a heavier tail. For $m_2 = 2$ and $m_1 = 1$ (Fig. 2, middle), the exponents of the fitted power-law distributions are 2.143 and 1.772 and $2/\phi_1 \approx 2.211$ and

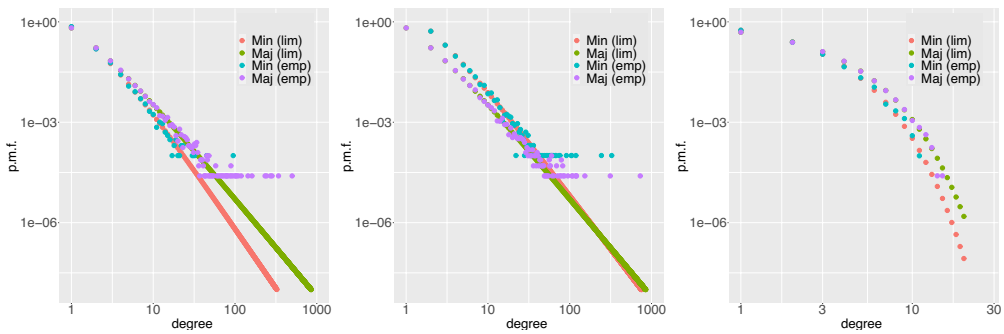


Fig. 2 Empirical and limiting degree distributions of homophily networks with 50,000 nodes, $\pi_1 = 0.2$, $\kappa_{11} = \kappa_{22} = 2$, $\kappa_{12} = \kappa_{21} = 1$: (left) $\alpha = 1$, $m_1 = m_2 = 1$; (middle) $\alpha = 1$, $m_1 = 2$, $m_2 = 1$; (right) $\alpha = 0$, $m_1 = m_2 = 1$

$2/\phi_2 \approx 1.920$. Finally, for the uniform attachment network (Fig. 2, right), the empirical and the limiting exponential tails of the distributions also show a good agreement.

3.2 Page-rank distribution

We recall first the definition of the Page-rank scores with damping factor c (Page et al. 1999). For attributed network model G_n, the Page-rank scores of nodes $v \in G_n$ with damping factor c is the stationary distribution $\{R_c(v) : v \in G_n\}$ of a random walk with jumps. At each step, with probability c, the walk follows an outgoing edge chosen uniformly at random among the possible available choices from the current node location in the network, while with probability $1-c$, it jumps to a uniformly selected node of the network. The Page-rank scores of the nodes are given as the solution to the linear system of equations:

$$R_c(v) = \frac{1-c}{|V|} + c \sum_{u \in V^-(v)} \frac{R_c(u)}{\deg^+(u)}, \quad v \in G_n \tag{4}$$

where $V^-(v)$ is the set of nodes with edges pointed to v and $\deg^+(u)$ is the out-degree of node u. At nodes with zero out-degree, the random walk stays in place with probability c and jumps to a uniformly chosen node with probability $1-c$. A high Page-rank value of a node results from the node either having a high in-degree or having an in-bound neighbor with a high Page-rank score.

Several asymptotic properties for the Page-rank scores of the linear dynamic attributed network model can be derived (Antunes et al. 2023a). It will be easier to describe the results in terms of the graph normalized Page-rank scores $\{\overline{R}_c(v) : v \in G_n\} := \{nR_c(v) : v \in G_n\}$ (Garavaglia et al. 2020). As n tends to infinity, the limiting distribution function of the (normalized) Page-rank scores per attribute has a power law tail with the same exponent $2/\lambda_c$ across all attributes, where λ_c can be explicitly computed [see Antunes et al. 2023a, Equation (4.5)]. This implies that for linear networks generated with this model, the *tail exponent of the limiting Page-rank scores distribution does not depend on the attribute type*, in contrast with the result on the asymptotic degree distribution in Sect. 3.1.

Additionally, if all rows of the matrix κ are identical, then it can be shown that in this case $\lambda_c = 1 + c$ for any $c \in (0,1)$ and $\phi_a = 1$ for all $a \in A$. This implies that the limiting Page-rank tail exponent is $2/(1+c)$ and the tail exponent of the limiting degree distribution is 2. In particular, these exponents are independent of the out-degree m_a of nodes. However, as in the case of only one attribute studied in Banerjee and Huang (2023), the out-degree significantly influences the degree separation between the "hubs" (maximal degree nodes) and the remaining nodes. Although the degree tail exponents are the same across attributes in this case, increasing the out-degree of a given type will lead to the maximal degree node coming from the same type with high probability. On the other hand, when $m_a = m \geq 1$ for all $a \in S$, the tail exponents for the limiting Page-rank distribution, as well as the limiting degree distribution, match in the tree ($m_a = 1$) and non-tree ($m_a > 1$) cases.

In spite of the degree exponent of the limiting the Page-rank distribution being insensitive to the attribute type, the bulk of the distribution depends on the attribute. This implies for instance, that the average Page-rank of nodes per attribute differs and can also be explicitly computed from the model as n tends to infinity [see Antunes et al. 2023a, Equation (4.6)].

In the case of uniform attachment model ($\alpha = 0$) with $m_a = 1$, $a \in S$ (which can also be extended to non-tree setting), the power-law tail result for Page-rank also holds but now with exponent $1/c$, while the degree distributions have exponential tails dependent on the attribute. This might appear surprising and seems to be new in the literature of network models. Intuitively, this can be understood by noting that as stated above, a high Page-rank value of a node results from having a high in-degree or an in-bound neighbor with a high Page-rank score. For the dynamic network discussed

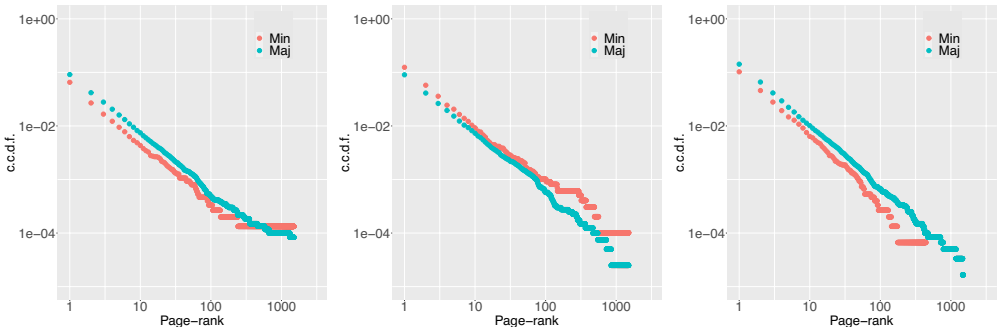


Fig. 3 Empirical complementary c.d.f. of the Page-rank distributions of homophily networks with 50,000 nodes, $\pi_1 = 0.2$, $\kappa_{11} = \kappa_{22} = 2$, $\kappa_{12} = \kappa_{21} = 1$: (left) $\alpha = 1$, $m_1 = m_2 = 1$; (middle) $\alpha = 1$, $m_1 = 2$, $m_2 = 1$; (right) $\alpha = 0$, $m_1 = m_2 = 1$

here, "older" nodes tend to have higher in-degrees and are typically close to other high degree (and high Page-rank) nodes. This reinforcement results in the Page-rank having heavier tails than degree.

Figure 3 depicts the empirical complementary cumulative distribution function (c.c.d.f.) of the Page-rank distributions for linear and uniform attachment networks with homophily using the same parameters as in Fig. 2 with damping factor $c = 0.85$. For $\alpha = 1$ and $m_1 = m_2 = 1$, the maximum likelihood estimates of the empirical tail exponents are 1.195 (minority) and 1.058 (majority) which are close to $2/\lambda_c \approx 1.079$. However, the average (normalized) Page-rank for minority and majority are 0.517 and 1.120, resp. With $m_1 = 2$, $m_1 = 1$, the exponents of the fitted power-law distributions are 1.114 (minority) and 1.110 (majority) and $2/\lambda_c \approx 1.047$. For uniform attachment network, we have 1.190 (minority) and 1.138 (majority) and $1/c \approx 1.176$.

4 Network sampling representation for rare minority

In this section, we consider the network model in a setting where there is a particular rare minority (attribute 1) which has higher propensity to connect within itself as opposed to majority nodes. In the context of network sampling, we devise schemes that increase the probability of sampling a minority node and therefore its representation in the sample.

4.1 Sampling methods

In the above setting, uniform and (total) degree-based sampling schemes are not efficient in sampling rare minorities if the sample size is much smaller than the network size and will be considered as baseline methods for comparison. Sampling methods that explore the network locally around the initial (uniformly sampled) random node by traversing edges along their natural direction have a higher efficacy for sampling rare minorities. We will propose such sampling schemes which are based on in-degree and Page-rank centrality measures.

Uniform sampling (*Unif*): sample a node uniformly at random from G_n.

Sampling proportional to degree (*Deg*): pick a node at random from the network and then sample a neighbor of this node uniformly at random.

Sampling proportional to in-degree (*InDeg*): select a node at random and then sample a node from one of its outgoing edges chosen at random. If the root node is picked (in a tree-network) then the root is sampled.

Sampling proportional to Page-rank with damping factor c (PR_c): pick a node uniformly at random from the network and then generate each time independently a geometric random variable X with parameter $(1 - c)$ with support starting at zero. Starting from the picked node, walk X steps at random using the directions of edges. The terminal node is sampled. If the root node is reached before X steps in a tree network, pick this node as the sampled node. Sampling a node with probability proportional to the Page-rank scores $\{R_{v,c}(n) : v \in G_n\}$ as defined in Sect. 3.2 is equivalent to the local algorithm PR_c in the context of the (tree) network model (Chebolu and Melsted 2008).

Fixed length walk sampling ($FixL_M$): Set $M \geq 0$. Consider the same implementation of the Page-rank scheme but now the number of walk steps taken is fixed and equal to M. Since $M = 0$ and $M = 1$ corresponds, respectively, to uniform sampling and sampling proportional to in-degree, we will consider $M \geq 2$.

4.2 Tree networks

We consider an asymmetric homophily scenario, where type 1 (minority) nodes are relatively rare compared to type 2 (majority) nodes and newly entering majority nodes have equal propensity to connect to minority or majority nodes. Minorities have relatively much higher propensity to connect to other minority nodes, as compared to majority nodes, namely,

$$\kappa_{11} = \kappa_{22} = \kappa_{12} = 1, \kappa_{21} = a, \qquad \pi_1 = \frac{\theta}{1+\theta}, \quad a, \theta \ll 1. \tag{5}$$

We analyze a linear ($\alpha = 1$) tree network of large size where a and θ are dependent, that is, $\theta = D\sqrt{a}$, where D is a positive constant. Let v be a node sampled from the network G_n and $a(v)$ its attribute, under the above sampling schemes. Table 2 summarizes our findings for the asymptotic probability of sampling a minority node under the above sampling schemes (the results are proved in the technical report Antunes et al. 2023a). We investigate how the relative performances of these schemes hold in a non-asymptotic regime for (sub-)linear tree and non-tree networks.

We generate a linear 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,

Table 2 Linear tree network: asymptotic sampling probabilities of a minority node as $n \to \infty$, $a \to 0^+$

Sampling	$\mathbb{P}(a(v) = 1 \mid G_n)$
Unif	$D\sqrt{a} + O(a)$
Deg	$2D\sqrt{a} - (4D^2 + \frac{1}{2})a + O(a^{3/2})$
InDeg	$3D\sqrt{a} + O(a)$
PR_c ($c \to 1$) and $FixL_M$	$\frac{(2D^2 - \frac{1}{2} + \sqrt{(2D^2-1/2)^2+4D^2})}{(2D^2+\frac{1}{2}+\sqrt{(2D^2-1/2)^2+4D^2})} + O(a)$

Table 3 Synthetic networks: structural properties (see Table 1 for the used notation)

	$\|V\|$	$\|E\|$	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	0.029	1.837	0.050	0.864	0.002	0.085	0.052	0.948
Syn. 2	10^5	99999	7.155	0.988	0.133	1.052	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

Table 4 Linear tree network (Syn. 1): estimated probability of sampling a minority node, and its average degree-rank and Page-rank in the network

Sampl. scheme	*Unif*	*Deg*	*InDeg*	$PR_{1/2}$	$PR_{2/3}$	$PR_{3/4}$	$PR_{4/5}$	$FixL_2$	$FixL_3$	$FixL_4$
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 5 Sub-linear tree network (Syn. 2): estimated probability of sampling a minority node, and its average degree-rank and Page-rank in the network

Sampl. scheme	*Unif*	*Deg*	*InDeg*	$PR_{2/3}$	$PR_{3/4}$	$PR_{4/5}$	$PR_{5/6}$	$FixL_4$	$FixL_5$	$FixL_6$
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 6 Linear non-tree network (Syn. 3): estimated probability of sampling a minority node, and its average degree-rank and Page-rank in the network

Sampl. scheme	*Unif*	*Deg*	*InDeg*	$PR_{1/2}$	$PR_{2/3}$	$PR_{3/4}$	$FixL_2$	$FixL_3$	$FixL_4$
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

$\pi_1 \approx 0.052$. The homophily and structural characteristics of the network are given in Table 3 (Syn. 1). Note that D_1 is large while D_2 is close to 1, and $H_{12}(< 1)$ is smaller than H_{21} corresponding to an asymmetric homophily. We estimate the probability of sampling a minority node (in each trial) for the sampling schemes defined above. We repeat the procedure to sample a node for each scheme 10^4 times and compute the proportion of minority nodes which were sampled. Additionally, we also compute the average of the degree-rank and Page-rank of the minority nodes sampled with respect to the whole network. The ranks are expressed as percent, where higher rank corresponds to smaller top percent. For example, a 5% result means that the minority nodes sampled are on average in top 5% of nodes with the highest degree (Page-rank) in the whole network. The results are given in Table 4. 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 proportional to Page-rank (PR_c) with $c = k/(k+1)$, $k \in \mathbb{N}$, the mean number of walk steps is k. 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 3 (Syn. 2). We estimate the probability of sampling a minority and its importance for each sampling scheme using 10^4 runs—see Table 5. 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.

4.3 Non-tree networks

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 3 (Syn. 3). As seen from the results (averaged over 10^4 runs) in Table 6, 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).

4.4 Discussion

In a setting where there is a small minority with higher propensity to connect within itself and majority nodes have equal preference to connect to any type of node, we argued that a sampling method that explores the network locally around a node selected at random followed by a fixed number of steps using the directions of edges has a higher probability to sample a minority node. It also finds more "popular" (higher degree and Page-rank) minority nodes. This is particularly more relevant for tree networks.

5 Centrality-based sampling and higher ranked attribute

The goal of this section is to investigate the role of the model parameters on the representation of the minority nodes when a fraction of nodes are sampled based on a centrality measure. The following schemes are considered:

A: sample (select) a fraction γ of nodes of the network with the highest centrality measure;
B: sample without replacement a fraction γ of nodes with probability proportional to the centrality measure.

The scheme *A* is a non-probabilistic sampling method where nodes are selected using a deterministic criterion. In this case, the nodes selected represent the top γ fraction of nodes with the highest centrality measure in the whole network. We quantify the proportion of nodes for each attribute type in both sampling schemes. If an attribute type dominates the other attribute type (i.e. with a proportion greater than 0.5) in a given sampling scheme, we call it the *higher ranked attribute* for that scheme.

5.1 Degree centrality

In this subsection, we consider the degree centrality of nodes in linear preferential attachment networks. We recall from Sect. 3.1 that η_a represents the limit of the normalized sum of the degrees and $2/\phi_a$ denotes the power law exponent of the limiting degree distribution of nodes with attribute a (see Eqs. 2 and 3). We can relate these quantities to conditions for the minorities to be higher ranked in the sampling schemes above for a *small* fraction γ, say top 1–3% which is the most interesting case. If the degree distribution tail of type 1 attribute is heavier than that of type 2 (i.e. $\phi_1 > \phi_2$), then type 1 has a higher proportion of nodes in the sample under scheme *A*. On the other hand, if $\eta_1 > \eta_2$, then a sampled node is more likely to be of type 1 [see the discussion following Eq. (2)] and therefore minority nodes are ranked higher under scheme *B*.

We consider below the three network scenarios of Sect. 2.3 using different configurations of the propensity matrix κ and give the conditions for minorities to rank higher in both sampling schemes as a function of the model parameters: node out-degrees (m_1, m_2) and node attribute probabilities (π_1, π_2). The proofs of these results are given in Antunes et al. (2023a).

5.1.1 Heterophily

We first consider a heterophilic network given by $\kappa_{11} = \kappa_{22} = 1$ and $\kappa_{12} = \kappa_{21} = K$, where $K \gg 1$. As K grows, ϕ_1 and ϕ_2 approach

$$\phi_1 \approx 2\left(1 - \frac{m_1\pi_1}{m_1\pi_1 + m_2\pi_2}\right), \qquad \phi_2 \approx 2\left(1 - \frac{m_2\pi_2}{m_1\pi_1 + m_2\pi_2}\right). \tag{6}$$

Thus, if $m_1\pi_1 < m_2\pi_2$, which always holds in the case $m_1 \le m_2$, then $\phi_1 > \phi_2$ and the minorities rank higher under scheme *A*, as intuitively can be understood by noting that the majority nodes boost up the minority ranks under scheme *A*. But, if $m_1\pi_1 > m_2\pi_2$, the minority nodes boost up the majority ranks under scheme *A* by connecting to them with more edges per incoming node.

As K becomes larger,

$$\eta_1 \approx \eta_2 \approx \frac{m_1\pi_1 + m_2\pi_2}{2}, \tag{7}$$

and thus the discrepancy in relative ranking between the two groups decreases under scheme *B*.

Figure 4 (left) shows the proportion of minority nodes for $0 < \gamma \le 1$ in the two sampling schemes when $m_1 \le m_2$—the network parameters are given in the caption of the figure. The structural properties of the network with $m_1 = m_2 = 2$ are given in Table 7 for the several considered scenarios. For small γ, the minority nodes rank higher under scheme *A* (the result for the majority is the complementary proportion). In this setting $\phi_1 < \phi_2$ and the empirical degree distribution of the minority is heavier than that of the majority which results in higher node degrees for the minority—see Fig. 5

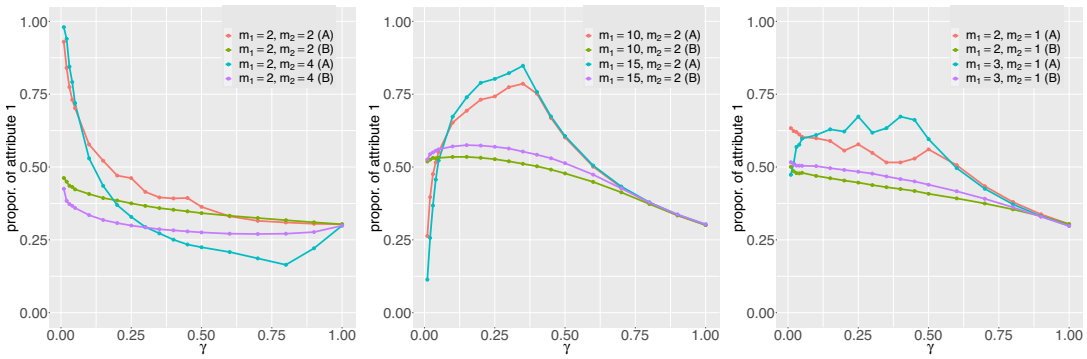


Fig. 4 Degree centrality: proportion of minority nodes under sampling schemes A and B of heterophilic networks with 30,000 nodes, $\alpha = 1$, $\pi_1 = 0.3, \kappa_{11} = \kappa_{22} = 1, \kappa_{12} = \kappa_{21} = 15$

Table 7 Synthetic network scenarios: structural properties with $\alpha = 1$, $\pi_1 = 0.3$, $m_1 = m_2 = 2$: heterophilic ($\kappa_{11} = \kappa_{22} = 1$, $\kappa_{12} = \kappa_{21} = 15$); homogenous homphily ($\kappa_{11} = \kappa_{22} = 15$, $\kappa_{12} = \kappa_{21} = 1$); asymmetric homophily ($\kappa_{11} = 15$, $\kappa_{22} = \kappa_{12} = \kappa_{21} = 1$). (See Table 1 for the description of remaining quantities.)

	$\lvert V\rvert$	$\lvert E\rvert$	D_1	D_2	H_{12}	H_{21}	$\frac{\lvert E_{11}\rvert}{\lvert E\rvert}$	$\frac{\lvert E_{22}\rvert}{\lvert E\rvert}$	$\frac{\lvert E_{12}\rvert}{\lvert E\rvert}$	$\frac{\lvert E_{21}\rvert}{\lvert E\rvert}$	$\frac{\lvert V_1\rvert}{\lvert V\rvert}$	$\frac{\lvert V_2\rvert}{\lvert V\rvert}$
Het	30^4	59997	0.198	0.097	1.347	3.087	0.018	0.047	0.284	0.651	0.302	0.698
Homo	30^4	59997	2.880	1.389	0.195	0.092	0.259	0.681	0.041	0.019	0.300	0.700
Asy. homo	30^4	59997	3.046	0.796	0.112	1.469	0.279	0.387	0.024	0.310	0.303	0.697

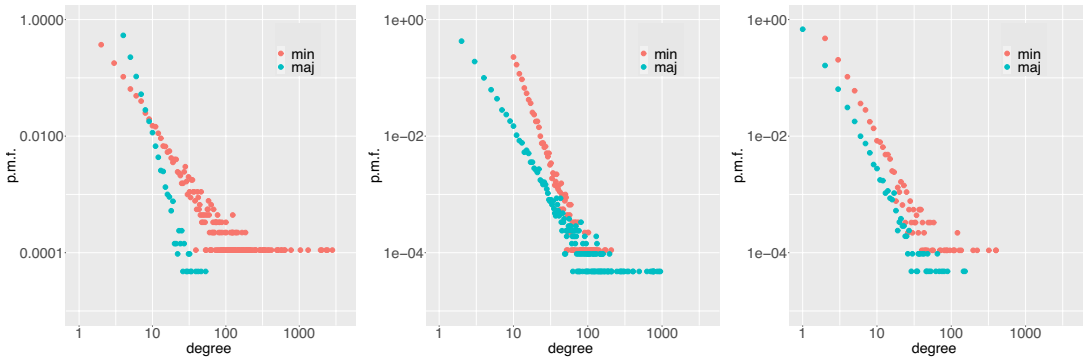


Fig. 5 Empirical degree distributions of networks with 30,000 nodes, $\alpha = 1$, $\pi_1 = 0.3$, $\kappa_{11} = \kappa_{22} = 1$, $\kappa_{12} = \kappa_{21} = 15$: (left) $m_1 = 2, m_2 = 4$, (middle) $m_1 = 10, m_2 = 2$, (right) $m_1 = 2, m_2 = 1$

(left) with $m_1 = 2, m_2 = 4$. Under scheme B, the empirical value η_1 is slightly smaller than η_2 which explains that the proportion of minority nodes in both settings is smaller than the majority for small γ. For other top ranks (say 10% and 20%) the minority is over-represented under both schemes in the sense that the proportion is higher than π_1, which is obviously approached when $\gamma = 1$.

Figure 4 (middle) depicts the case where $m_1\pi_1 > m_2\pi_2$. For small γ, the majority nodes rank higher under scheme A. The degree distribution of the majority in Fig. 5 (middle) with $m_1 = 10, m_2 = 2$ is heavier ($\phi_1 > \phi_2$) and thus the majority nodes have higher degrees. On the other hand, for small γ, the proportion of minority nodes is close but above 0.5 under scheme B (the empirical value η_1 is slightly higher that η_2). For other values $\gamma = 0.1, 0.2, 0.3$, the minority nodes rank higher in both schemes. This is due to the fact the lower bound of the degree of minority nodes m_1 is large and the heterophilic scenario.

Figure 4 (right) represents the case where ϕ_1 and ϕ_2 are close. The degree distributions are given in Fig. 5 (right) with $m_1 = 2$, $m_2 = 1$, where ϕ_1 is slightly larger than ϕ_2 and thus the minority rank higher under scheme A for small γ (when $m_1 = 3$, $m_2 = 1$ it is the opposite). There is almost no discrepancy in relative ranking between the two groups for small γ under scheme B.

5.1.2 Homogenous homophily and homogeneous mixing

When the network is homogeneous mixing, that is, $\kappa_{ij} = 1$, for all $i,j = 1,2$, or when there is strong homophily $\kappa_{11} = \kappa_{22} = K \gg 1$ and $\kappa_{12} = \kappa_{21} = 1$, we have

$$\phi_1 \approx 1, \qquad \phi_2 \approx 1$$

and

$$\eta_1 \approx m_1\pi_1, \qquad \eta_2 \approx m_2\pi_2,$$

as $K \to \infty$. Thus, if $m_1\pi_1 > m_2\pi_2$, the minority nodes rank higher via scheme B. Moreover, although the degree tail exponents are comparable, if m_1/m_2 is large enough, the degrees of minority nodes get a high initial boost. The analysis of Banerjee and Bhamidi (2021), Galashin (2016), extended to the multi-attribute setting, suggests a "persistence phenomenon", namely, the maximal degree nodes from any attribute type emerge from, with high probability, the oldest nodes of that type added to the network. Consequently, minority nodes also seem to have a higher ranking under scheme A, when γ is small. Thus, in the context of social networks, increasing the "social interaction" (quantified by m_1/m_2) for the minority nodes increases their popularity under both schemes in this setup.

Figure 6 (left) considers homogeneous homophily networks, where m_1 is sufficiently larger than m_2 such that the minority rank higher in both sampling schemes for small γ. This also true for other top 10% and 30% when m_1 is large since it determines the lower bound of the degree of the minority nodes as noted above. On the other hand, in Fig. 6 (middle) the ratio $m_1/m_2 = 1.5$ is not sufficiently large for the minority nodes to rank higher under schemes A and $m_1\pi_1 < m_2\pi_2$ in scheme B. With the same outgoing edges ($m_1 = m_2 = 2$), the proportion of nodes from attribute 1 with both scheme remains approximately close to π_1. Figure 6 (right) considers two homogeneous mixing networks where the same conditions for the minority nodes to rank higher hold.

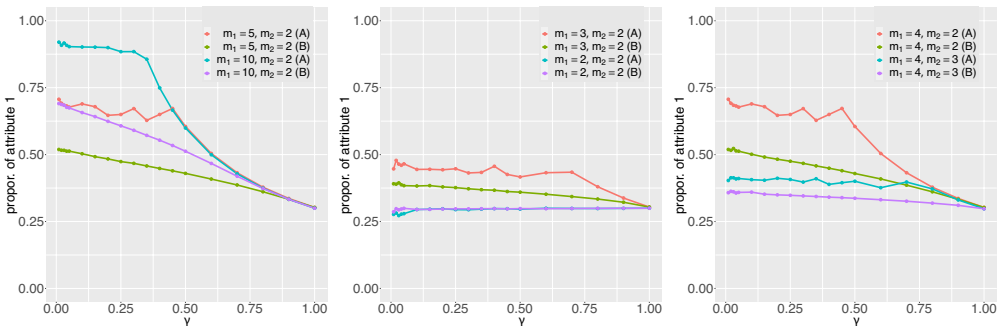
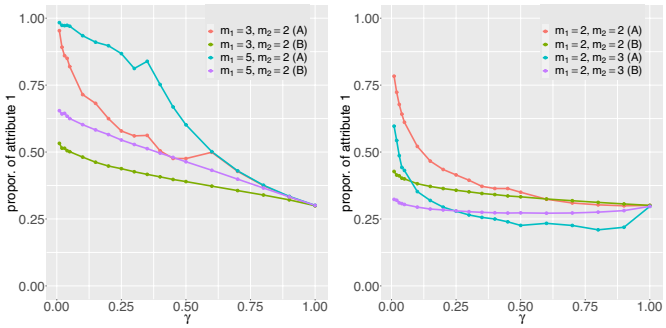


Fig. 6 Degree centrality: Proportion of minority nodes under sampling schemes A and B of homogeneous homophily and homogeneous mixing networks with 30,000 nodes, $\alpha = 1$, $\pi_1 = 0.3$, $\kappa_{11} = \kappa_{22} = 15$, $\kappa_{12} = \kappa_{21} = 1$ (left and middle) and $\kappa_{11} = \kappa_{22} = \kappa_{12} = \kappa_{21} = 1$ (right)

Fig. 7 Degree centrality: Proportion of minority nodes under sampling schemes A and B of asymmetric homophily networks with 30,000 nodes, $\alpha = 1$, $\pi_1 = 0.3$, $\kappa_{11} = 15$, $\kappa_{12} = \kappa_{12} = \kappa_{21} = 1$



5.1.3 Asymmetric homophily

Finally, we consider the strong asymmetric homophily regime, namely, $\kappa_{11} = K \gg 1$, $\kappa_{22} = \kappa_{12} = \kappa_{21} = 1$. As $K \to \infty$,

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

and

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

Thus, since $\phi_1 > \phi_2$, the minorities rank higher under scheme A. In comparison, for scheme B, minorities rank higher if $2m_1\pi_1 > m_2\pi_2$. Again, in the context of social networks, this implies that, if the majority nodes do not show appreciable attribute bias when connecting to the network, the minorities can increase their popularity by enhancing their connectivity preference towards other minority nodes. The last two scenarios under scheme A were briefly considered heuristically in Espín-Noboa et al. (2022) using fluid limits.

In Fig. 7 (left) the condition $2m_1\pi_1 > m_2\pi_2$ holds and the minority nodes dominates under scheme B for small γ. With $m_1 = 5, m_2 = 2$, the minority node can rank higher for top 20% and 30%. Figure 7 (right) illustrates the case when $2m_1\pi_1 < m_2\pi_2$, however, with $m_1 = m_2 = 2$ the minority can be over-represented for $\gamma \in (0, 1)$. On the other hand, under scheme A minority always dominates in all the cases considered for small γ and also for larger top-ranks if $m_1 > m_2$.

5.1.4 Discussion

In the heterophily regime considered, if we select a small fraction γ (say, $\gamma \leq 0.03$) of nodes of the network with the highest degree (scheme A), the proportion of minority nodes selected is higher (rank higher) if the number of outgoing edges of a minority node m_1 is smaller than or equal to the number outgoing edges of a majority node m_2. On the other hand, if nodes are sampled proportional to their degrees (scheme B), the proportions of the minority and majority are similar. With homogenous homophily, the minority nodes rank higher under scheme A if the ratio m_1/m_2 is sufficiently larger. However, under scheme B this depends also on the proportion of the minority nodes in the network π_1 ($m_1\pi_1 > m_2\pi_2$). In the case of asymmetric homophily the condition for the minority to have a larger proportion in the sample under scheme B is less restrictive ($2m_1\pi_1 > m_2\pi_2$), while the minority always ranks higher under scheme A.

5.2 Page-rank centrality

In this subsection, we consider the Page-rank centrality measure in the schemes A and B, and explore the proportion of minority nodes in the sample for the heterophily and homophily scenarios considered in Sect. 5.1. As for the degree centrality, if the normalized sum of the Page-ranks of the minorities (majorities, resp.) is higher than the majorities (minorities, resp.), the probability of sampling a minority (majority, resp.) node in each draw is higher and hence the minority (majority, resp.) rank higher in scheme B for small γ. On the other hand, under scheme A, Page-rank distribution tails for the two attribute types are expected to have the same power-law exponents (see Sect. 3.2). However, one distribution tail can still dominate the other as, for example, in Fig. 3 or homogenous homophily/mixing in Section 5.1. As in the latter section, we expect this to depend on the values of m_1 and m_2. The derivation of the conditions for the minority to rank higher in terms of the model parameters seems theoretically challenging, and we explore the issues numerically to gain insight.

5.2.1 Heterophily

We consider a heterophilic linear network with $\kappa_{11} = \kappa_{22} = 1$ and $\kappa_{12} = \kappa_{21} = 15$ and $\pi_1 = 0.3$ as above for comparison and set the damping factor to $c = 0.85$. Figure 8 (left) shows two settings with $m_1 \leq m_2$, where there is almost no discrepancy in relative ranking between the minority and majority

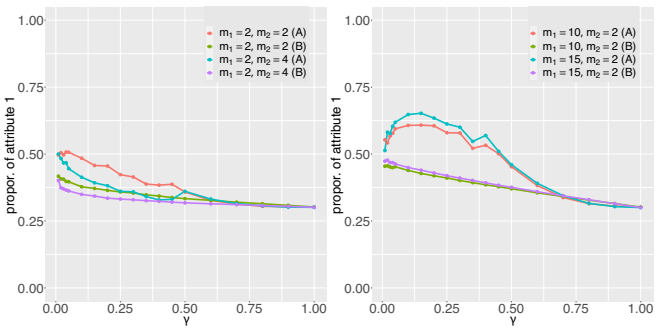


Fig. 8 Page-rank centrality: proportion of minority nodes under sampling schemes A and B of heterophilic networks with 30,000 nodes, $\alpha = 1$, $\pi_1 = 0.3$, $\kappa_{11} = \kappa_{22} = 1, \kappa_{12} = \kappa_{21} = 15$

for small γ (top 1–3%) under scheme A. This contrasts with the situation for the degree centrality measure in Fig. 4 (left). As the Page-rank of majority nodes is increased by high Page-rank minority nodes that connect to them, the separation between the proportions of minority and majority nodes in the sample chosen according to scheme A decreases.

The lower proportion of attribute 1 (minority) with $m_1 = 2, m_2 = 4$ compared to $m_1 = 2, m_2 = 2$ under scheme A (when γ is not too large) can be explained as follows. When $m_1 < m_2$, the Page-rank formula (4) shows that Page-rank of minority nodes is diminished. To see this, observe that, due to heterophily, for a minority node most of its inbound neighbors belong to the majority. From the Page-rank formula, the contribution to the Page-rank of a minority node by a majority node u is $O(\frac{1}{\deg^+(u)} = \frac{1}{m_2})$, which is small as m_2 increases. This reduces the proportion of minority nodes selected by scheme A, as seen in Fig. 8 (left). In both settings, under scheme B, the normalized sum of the Page-ranks is smaller for the minorities which explain that the majority rank higher for small γ.

Figure 8 (right) depicts two cases with $m_1 > m_2$ where the minority can rank higher under scheme A for larger top-ranks γ (20% and 30%) by a reasoning similar to above. Now, the minority nodes have larger out-degrees, and their contribution to the ranks of the outbound majority neighbors is largely diminished ($O(1/m_1)$) and the minority ranks higher. We also see that under scheme B, the proportion of minority increases.

5.2.2 Homogeneous homophily

We set the propensity matrix to $\kappa_{11} = \kappa_{22} = 15$ and $\kappa_{21} = \kappa_{12} = 1, \pi_1 = 0.3$ and $c = 0.85$. With $m_1 = m_2 = 2$, the proportion of minority nodes in both schemes is close to its expected proportion π_1 in the network under both schemes and this proportion is slightly higher with $m_1 = 5, m_2 = 2$ for small γ – see Fig. 9 (left). As seen from Fig. 9 (right), for the minority to rank higher, it is needed that the number of minority nodes they connect to is large under scheme A, which roughly amounts to $m_1 \gg m_2$ in the homophilic regime. As found in Banerjee and Huang (2023) for only one attribute, the out-degree significantly influences the degree separation between the "hubs" (maximal degree nodes) and the remaining nodes. Increasing the out-degree of a given type will also lead to the maximal degree node coming from the same type with high probability. Additionally, for the dynamic networks considered here, "older" nodes tend to have higher in-degrees and are typically close to other high degree (and high Page-rank) nodes. Putting it all together and by noting that a high Page-rank value of a node results from the node having either a high in-degree or having an in-bound neighbor with a high Page-rank score, reinforces the conditions for the minority to rank higher for small γ under scheme A with homophily.

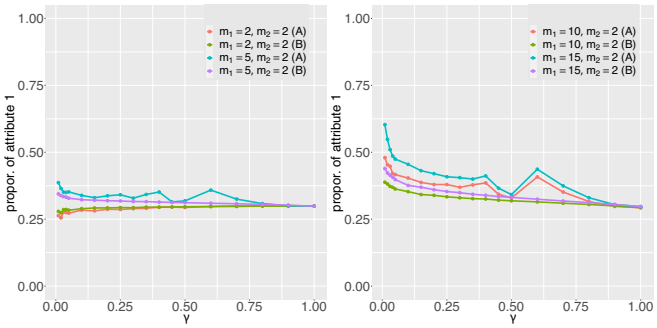


Fig. 9 Page-rank centrality: proportion of minority nodes under sampling schemes A and B (Page Rank) of homogenous homophily networks with 30,000 nodes, $\alpha = 1, \pi_1 = 0.3$, $\kappa_{11} = \kappa_{22} = 15$

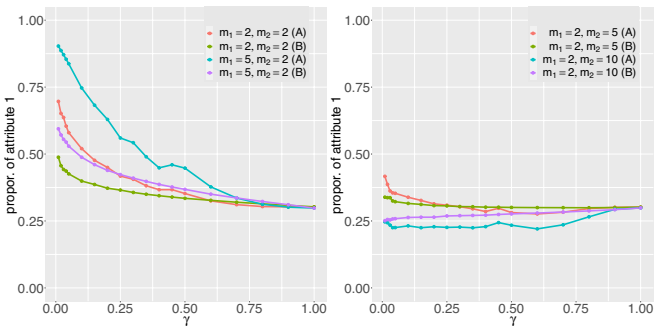


Fig. 10 Page-rank centrality: proportion of minority nodes under sampling schemes A and B of asymmetric homophily networks with 30,000 nodes, $\alpha = 1, \pi_1 = 0.3, \kappa_{11} = 15$, $\kappa_{22} = \kappa_{12} = \kappa_{21} = 1$

5.2.3 Asymmetric homophily

In this scenario, the minority is homophilic with propensity matrix given by $\kappa_{11} = 15, \kappa_{22} = \kappa_{12} = \kappa_{21} = 1, \pi_1 = 0.3$ and $c = 0.85$. For the setting in Fig. 10 (left) where $m_1 \geq m_2$, the minority ranks higher under scheme A not only for small γ. For scheme B, the minority needs to increase their popularity through the number of outgoing edges to rank higher. On the other hand, if the majority nodes increase their out-degree even if they do not show appreciable attribute bias when connecting to the network and the minority is homophilic, the majority can rank higher in both schemes (Fig. 10 (right)).

5.2.4 Discussion

The results show that with Page-rank centrality measure in the heterophily regime under scheme A with γ small, the proportions of the minority and majority selected tend to be similar if their out-degrees m_1 and m_2 are equal, while under scheme B the majority ranks higher. Increasing m_1, the minority can rank higher for a large range of γ under scheme A, which also increases the visibility of minority nodes for scheme B. In a homogeneous homophily network, for the minority to rank higher under scheme A, the out-degree of minority has to be much larger than the out-degree of majority, which also increases the proportion of minority nodes sampled with scheme B. With asymmetric homophily, if $m_1 \geq m_2$, the minority ranks higher under scheme A for a large range of γ and in scheme B when m_1 is sufficiently greater than m_2.

6 Bias in ranking through degree centrality and subgraph sampling

In the context of the use of synthetic models for providing insight into real world systems, a different direction is the study of the performance of subgraph sampling methods in the representation and ranking of various attributes, especially in the tail of the distribution. More precisely, one has a partial observation of the nodes (or edges) in subgraph sampling and the goal is to infer the bias of the induced subgraph, especially tail properties like the connectivity structure of minority high-degree nodes from this partial measurement.

Fix a parameter p representing the density of items (nodes or edges) sampled from the network. The two main sampling schemes considered in this section are:

1. *Induced subgraph sampling:* Here one samples a proportion p of the nodes of the graph uniformly at random and observes the induced subgraph on the sampled nodes.
2. *Incident subgraph sampling:* Here one samples a proportion p of the edges of the graph uniformly at random and observes the induced subgraph generated by these edges.

The first type of sampling is representative in the construction of contact networks in social network research, when a sample of individuals of different types is first selected and then individuals are interviewed regarding some measure of contact among themselves (e.g. friendship, likes or dislikes, etc.). The second design is, for example, implicit in the construction of streaming graphs (e.g. Twitter) with different groups in a very large network, wherein edges (tweets) are sampled from the stream of edges, after which sender and the receiver nodes (users) are observed. Note that under incident subgraph sampling, the probability that a node is selected depends on its degree.

Write G_n^{ind} (resp., G_n^{inc}) for the induced (resp., incident) sampled graph from G_n. In either case, inferences about the underlying graph are then based on the subgraph. Now fix $0 < \gamma \leq 1$. The goal is to understand the composition of the top γ percentile of nodes as measured according to their degree distribution, and comparing the information provided by the sampled graph in contrast to the underlying network. For any attributed network G_n, rank the nodes in order of their degrees as in scheme A (Sect. 5). For quantile level γ, let $prop(G_n;\gamma)$ denote the proportion of type 1 (minority) nodes amongst the top γ proportion of nodes in terms of degrees. Now if $G_n^{\texttt{sample}} = G_n^{ind}$ or G_n^{inc} is a graph obtained by sampling from G_n using the schemes above, define the bias for the sampling scheme for quantile level γ as

$$bias(\texttt{sample};\gamma) = prop(G_n^{\texttt{sample}};\gamma) - prop(G_n;\gamma). \tag{10}$$

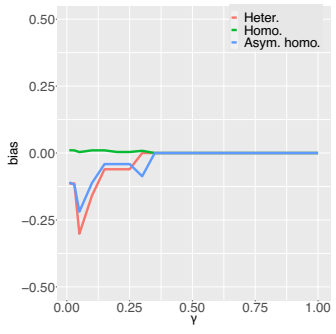


Fig. 11 Bias of induced subgraph sampling of linear tree networks for $\pi_1 = 0.3$, $p = 0.2$ with heterophily ($\kappa_{11} = \kappa_{22} = 1$, $\kappa_{12} = \kappa_{21} = 15$), homophily ($\kappa_{11} = \kappa_{22} = 15$, $\kappa_{12} = \kappa_{21} = 1$) and asymmetric homophily ($\kappa_{11} = 15, \kappa_{22} = 15, \kappa_{12} = \kappa_{21} = 1$)

6.1 Tree networks

To state an analytical result for the sampling bias under induced subgraph sampling for a tree network, we need some notation. Recall the limiting degree distribution according to attribute types in Sect. 3.1 and denote by D^a a random variable with the distribution $p_a(\cdot)$ for $a \in \{1,2\}$. For $p \in [0,1]$, write $\mathsf{Bin}(D^a,p)$ for a binomial random variable with D^a number of trials (conditionally on D^a generated first) and success probability p. Recall that the $(1-\gamma)$ percentile of the distribution of a random variable Z is given by the unique z such that $\mathbb{P}(Z<z) \le 1-\gamma$ and $\mathbb{P}(Z \le z) \ge 1-\gamma$. Let k_γ denote the $(1-\gamma)$ percentile of the distribution $\pi_1\mathbb{P}(D^1 \in \cdot) + \pi_2\mathbb{P}(D^2 \in \cdot)$. Write $\tilde{\gamma} = \pi_1\mathbb{P}(D^1 \ge k_\gamma) + \pi_2\mathbb{P}(D^2 \ge k_\gamma)$, noting that $\tilde{\gamma}$ might not equal γ owing to discretization effects. Similarly let $k_{\gamma,p}$ denote the corresponding percentile but for the distribution

$$\mu_p(\cdot) := \pi_1\mathbb{P}(\mathsf{Bin}(D^1,p) \in \cdot) + \pi_2\mathbb{P}(\mathsf{Bin}(D^2,p) \in \cdot),$$

and let $\tilde{\gamma}_p = \mu_p([k_{\gamma,p},\infty))$. Then, under appropriate assumptions, one obtains the following result for the induced subgraph sampling (the proof is given in Antunes et al. 2023a), as $n \to \infty$,

$$bias(\mathtt{sample};\gamma) \xrightarrow{\mathrm{P}} \pi_1\left(\frac{\mathbb{P}(\mathsf{Bin}(D^1,p) \ge k_{\gamma,p})}{\tilde{\gamma}_p} - \frac{\mathbb{P}(D^1 \ge k_\gamma)}{\tilde{\gamma}}\right). \tag{11}$$

The asymptotic bias given by (11) is plotted in Fig. 11 for the three network configurations of Sect. 2.3. For heterophilic and asymmetric homophily networks, the bias is negative for initial top ranks γ and approximately zero for homogeneous networks. The reasons for these results and the case when $m_1, m_2 > 1$ are investigated numerically in the next section.

6.2 Non-tree networks

In applications, a key issue of interest is over- or under-representation of minorities with subgraph samplings. Having a network model, it is interesting to understand the dependence of representation on the driving parameters of the model. We explore these questions below for the various homophily and heterophily scenarios considered in Sect. 5.

6.2.1 Heterophily

We consider the same propensity matrix as in Sect. 5 for this scenario ($\kappa_{11} = \kappa_{22} = 1, \kappa_{12} = \kappa_{21} = 15$) and the linear attachment model with $\alpha = 1, n = 30000, \pi_1 = 0.3$, but varying m_1 and m_2. For induced subgraph sampling, the densities p of sampled nodes are 0.1 and 0.2. To be consistent, we successively sample edges under incident subgraph sampling until the same densities of nodes are selected.

Figure 12 (left) depicts the bias for both subgraph sampling schemes when the number of outgoing edges for minority and majority nodes is equal to 2. Recall under this setting the proportion of type 1 nodes amongst the top γ fraction of nodes with the highest degree from Fig. 4 (left) - scheme A. In this case, for small γ, the minor ranks higher. Since nodes are sampled at random under induced subgraph sampling and p is small, sampling fails to capture the minority "hubs" (large degree nodes) and the proportion of type 1 nodes amongst the top γ degree nodes is smaller in the subgraph and the bias is negative. As γ approaches 1, it is expected that the bias is zero under induced subgraph sampling. The result is in line with the negative bias under induced subgraph sampling for tree networks (Fig. 11). In incident subgraph sampling while edges are in the sample with equal probability, the nodes are included with unequal probabilities that depend on the degree (minority node

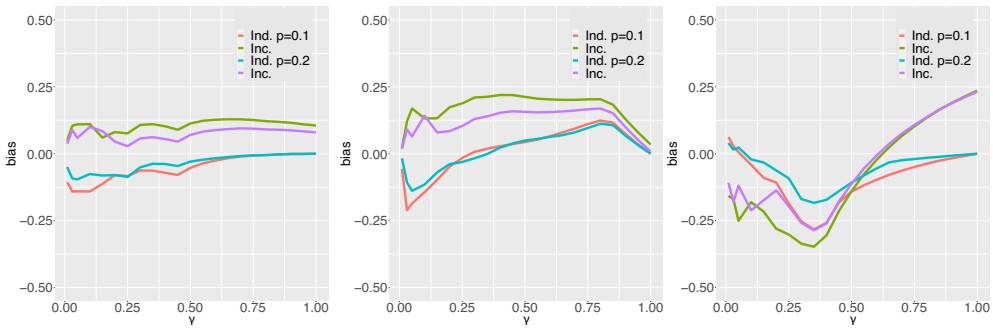


Fig. 12 Bias of induced and incident subgraph sampling of heterophilic networks with 30,000 nodes, $\alpha = 1$, $\pi_1 = 0.3$, $\kappa_{11} = \kappa_{22} = 1$, $\kappa_{12} = \kappa_{21} = 15$: (left) $m_1 = m_2 = 2$; (middle) $m_1 = 2, m_2 = 4$; (right) $m_1 = 10$ and $m_2 = 2$

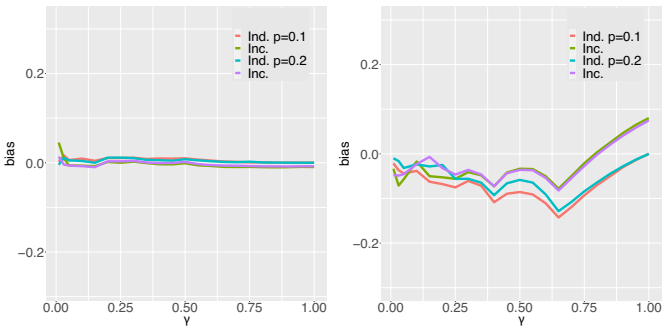


Fig. 13 Bias of induced and incident subgraph sampling of homogeneous homophily networks with 30,000 nodes, $\alpha = 1$, $\pi_1 = 0.3$, $\kappa_{11} = \kappa_{22} = 15$, $\kappa_{12} = \kappa_{21} = 1$: (left) $m_2 = m_2 = 2$; (right) $m_1 = 3$, $m_2 = 2$

"hubs" are more likely to be sampled). This explains that under the heterophilic scenario (with $m_1 = m_2 = 2$) the bias is positive.

Figure 12 (middle) shows the case $m_1 = 2$ and $m_2 = 4$. In this setting there are more minority nodes with larger (in-) degrees of connections from the majority nodes (cf. Fig. 5 (left)) which increases the bias for both subgraph samplings (say $0.01 < \gamma < 0.15$).

If the number of outgoing edges of minority increases to $m_1 = 10$ and $m_2 = 2$, the majority nodes have larger (in-) degrees in comparison to the case $m_1 \approx m_2$ (cf. Fig. 5 (middle)) and rank higher (cf. Fig. 4 (middle)) for small γ. Thus, incident subgraph sampling for small γ now shows a negative bias in Fig. 12 (right) since it is more likely to sample edges from minority nodes to majority nodes. However, for induced subgraph sampling, the sampled minority nodes have now more connections toward sampled majority nodes and can be over represented in the subgraph (positive bias) for small γ.

6.2.2 Homogeneous homophily

In this scenario ($\kappa_{11} = \kappa_{22} = 15$, $\kappa_{21} = \kappa_{12} = 1$), the bias is close to zero for both sampling methods when $m_1 = m_2 = 2$ – see Fig. 13 (left). We recall that the minority proportion is approximately π_1 in the original network for all γ (cf. Fig. 6 (middle)). The bias result also agrees with the induced subgraph sampling for tree networks (Fig. 11). The strong homophily reduces the effect of one type on the degree distribution of the other. Thus, the majority and minority nodes are seen in roughly equal proportions in the top γ percentiles under both sampling schemes as well as in G_n.

When the number of outgoing edges of minority nodes increases to $m_1 = 3$ and $m_2 = 2$, the proportion of minority nodes in the original network are over-represented and decreases almost linearly as γ increases (cf. Fig. 6, scheme A (middle)). However, it creates a negative bias for induced and incident subgraph sampling when γ is not close to 1 – see Fig. 13 (right). Other settings such as $m_1 = 5, m_2 = 2$ and $m_1 = 10, m_2 = 2$ where the minority rank higher (cf. Fig. 6 (left)) have also shown a negative bias and we omit the plots.

The practical recommendation is that minority should have the same out-degree as the majority to maintain their representation in the subgraphs.

6.2.3 Asymmetric homophily

Finally, consider the scenario where the minority is homophilic ($\kappa_{11} = 15, \kappa_{21} = 1$) and the majority has equal propensity to connect to any node in the network ($\kappa_{12} = \kappa_{22} = 1$). The bias is negative and positive under induced and incident subgraph samplings, respectively, for $m_1 = m_2 = 2$ (Fig. 14

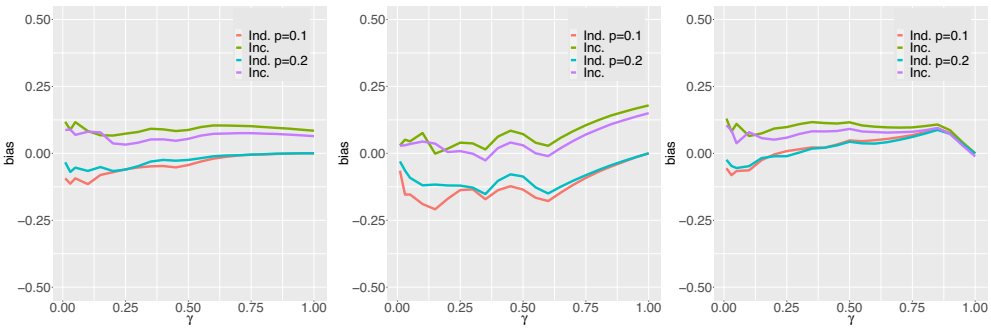


Fig. 14 Bias of induced and incident subgraph sampling of asymmetric homophily networks with 30,000 nodes, $\alpha = 1$, $\pi_1 = 0.3$, $\kappa_{11} = 15$, $\kappa_{12} = \kappa_{12} = \kappa_{21} = 1$: (left) $m_2 = m_2 = 2$; (middle) $m_1 = 3$, $m_2 = 2$; (right) $m_1 = 2$ and $m_2 = 3$

(left)). In this setting, the minority ranks higher only for small γ (cf. Fig. 7, scheme A (right)). This indicates that, in this regime, although the maximal degree nodes appear to be from the minorities, these are relatively few in number. Thus, these have very small chance of being sampled via induced subgraph sampling, which explains the negative bias. However, in incident subgraph sampling, the asymmetric homophily increases the probability of these being sampled (as there are many edges between minority nodes) and, thereby, creates a positive bias. The composition of the sample obtained via induced subgraph sampling are sensitive to small perturbations. Indeed, slightly changing m_1/m_2 from 1 leads to changes in bias (Fig. 14 middle and right).

6.2.4 Discussion

The results show that in the heterophily regime, if $m_1 \leq m_2$, the bias of induced (incident) subgraph sampling in the top percentile of high degree nodes is negative (positive). With homogeneous homophily the bias is close to zero for both subgraph samplings when $m_1 = m_2$. Finally, for asymmetric homophily the signs of the bias for induced and incident subgraph samplings are the same as in heterophily regime, if m_1 are m_2 are close.

7 Real-world networks

In this section, we provide evidence of the characteristics of the considered model and the insights of ranking of minorities using sampling in real networks. We analyze four publicly available datasets of real attributed networks from different domains and different homophily levels. Table 8 shows the network statistics of interest. Hate is a retweet network where nodes denote users, and edges represent retweets among them. Users in the dataset are classified as either "hateful" (minority) or "normal" (majority) depending on the sentiment of their tweets. The network is directed with asymmetric homophily where minority nodes have a higher propensity to connect to other minority nodes. APS is a scientific (directed) network from the American Physical Society where nodes represent articles from two subfields and

Table 8 Real-world networks: characteristics (see Table 1 for the used notation)

	$\|V\|$	$\|E\|$	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	4971	10170	26.621	0.519	1.204	1.565	0.318	0.412	0.117	0.153	0.109	0.891
APS	1853	3638	2.088	1.667	0.116	0.122	0.294	0.650	0.027	0.029	0.376	0.624
Wikipedia	2132	3143	1.695	1.081	0.693	0.737	0.040	0.774	0.090	0.096	0.153	0.847
Escort	16730	39044	0	0	2.090	2.090	0	0	39044	39044	0.396	0.604

Table 9 Network sampling for rare minority: Hate network (estimated probability of sampling a minority node, and its average degree-rank and Page-rank in the network)

Sampl. scheme	*Unif*	*Deg*	*InDeg*	$PR_{1/2}$	$PR_{2/3}$	$PR_{3/4}$	$FixL_2$	$FixL_3$	$FixL_4$
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

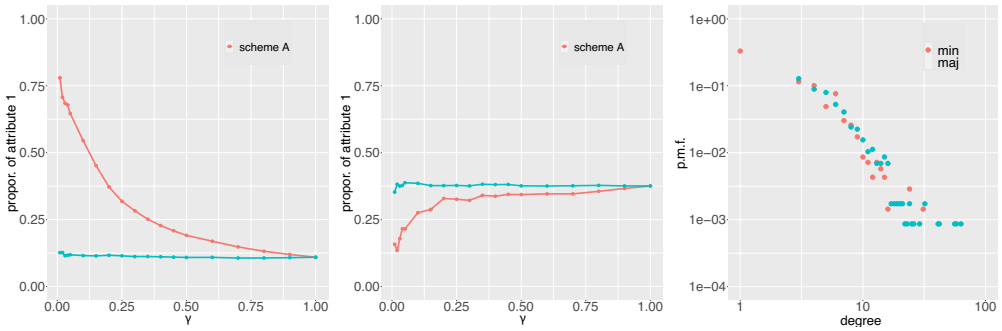


Fig. 15 Degree centrality: Proportion of minority nodes under sampling schemes A and B for Hate (left) and APS (middle) networks. Empirical degree distributions of APS network (right)

edges represent citations with a high homogeneous homophily. The Wikipedia dataset is a hyperlink (directed) network where nodes represent U.S. politicians with attributes as either male (majority) or female (minority) with a moderate homogeneous homophily. The Escorts dataset represents a (undirected) network of sexual contacts from Brazil. Nodes are of two types: client (majority) or escort (minority) exhibiting extreme heterophily.

7.1 Network sampling for rare minority

We inspect the Hate network which shares similar homophily characteristics with the synthetic networks considered in Sect. 4 (cf. Table 3) to assess the probability of sampling a minority node. We consider the largest connected component of the network (with diameter 24) for a fair comparison with the results provided in Table 4. For the several sampling schemes proposed, the results (averaged over 10^4 runs) in Table 9 are in line with the ones obtained with the 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 the network, where the proportions of edges from "hateful" to "normal" users is higher than in the synthetic network. This can also be seen from the homophily measures H_{12}.

7.2 Centrality-based sampling and higher ranked attribute

We consider the Hate and APS networks with power-law degree distributions to assess the ranking of the minorities under schemes A and B based on the degree (Sect. 5.1). For the Hate network, the exponents of the fitted degree distributions are 1.138 (minority) and 1.597 (majority). Figure 15 (left) shows that under scheme A, the minorities rank higher (since the degree distribution is more heavy-tailed) for small γ. Under scheme B, the minority do not rank higher due to the smaller normalized sum of the degrees (0.633 (minority) and 1.331 (majority)) but can maintain its rank in the sample. For the APS network, the minority ranks lower in both schemes – Fig. 15 (middle). The degree distributions are plotted in Fig. 15 (right) where the majority has a heavier-tailed distribution and thus ranks higher under scheme A. The normalized sum of the degrees is also larger for the majority in scheme B.

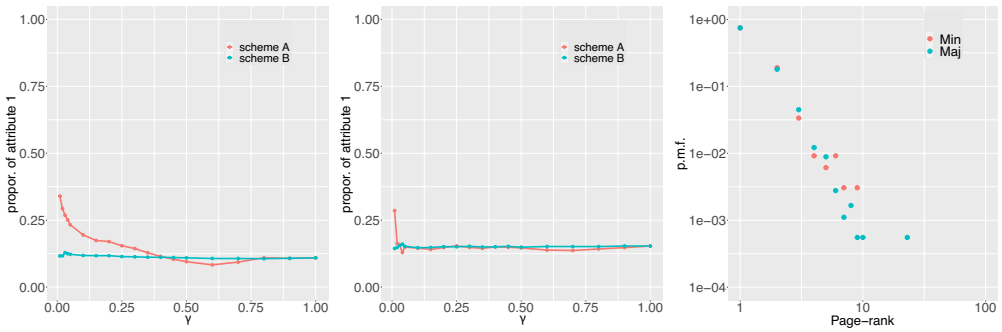


Fig. 16 Page-rank centrality: Proportion of minority nodes under sampling schemes A and B for Hate (left) and APS (Wikipedia) networks. Empirical Page-rank distributions of Wikipedia network (right)

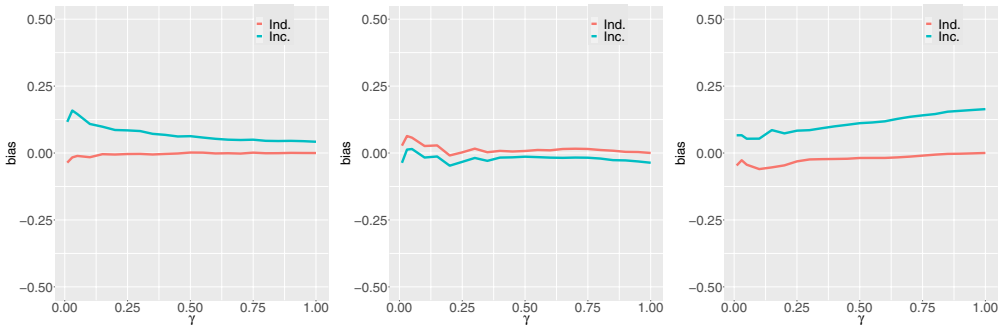


Fig. 17 Bias of induced and incident subgraph sampling of (left) Escort; (middle)APS; (right) Hate

We also consider the Page-rank centrality measure in the sampling schemes A and B (Sect. 5.2), and explore the relative ranking of the minority for the Hate and Wikipedia networks. The results are in line with those for the synthetic networks (asymmetric and homogeneous homophily), where the proportion of minority for a small fraction γ of nodes is higher under scheme A in the asymmetric case and equals the group size proportion with scheme B. Figure 16 shows the normalized Page-rank distributions for the Wikipedia network which have similar tail exponents for the two attributes (2.970 (minority) and 2.876 (majority)).

7.3 Bias of subgraph sampling for ranking through degree centrality

The over- or under-representation of minorities via induced and incident subgraph is given in Fig. 17. The signs of the bias for the Escort (heterophily), APS (homogeneous homophily) and Hate (asymmetric homophily) networks agree with the network model (Sect. 6.2).

7.4 Discussion

The findings for the three related problems investigated using an attributed network model are highly relevant for the considered real-world networks with different levels of homophily and heterophily.

8 Conclusions and future work

We investigated three related problems concerning sampling minorities in attributed networks. Using a dynamic attributed network model with homophily/heterophily, we provided analytical and numerical results in the representation, ranking and bias of minorities based on the degree and/or Page-rank centrality measures for several sampling schemes. We explained through the model parameters the under- and over-representation of minority nodes in the sample which can differ significantly from the original network. We also discussed how minorities can preserve their "position" in the sample. The findings and insights from the sampling analysis were assessed with real-world networks.

8.1 Limitations and future work

This paper has only considered a specific setting of nodal attribute models (directed networks, two attributes) and there are research questions that still need to be explored. A partial list includes:

1. More detailed understanding of the sublinear regime, both analytically and through numerics. The model without attributes exhibits fascinating degree distributional asymptotics, and for questions such as seed detection and network archaeology, also exhibits phase transition at $\alpha = 1/2$ (Banerjee and Bhamidi 2021).
2. In the setting of the linear $\alpha = 1$ regime, while the tail exponent of the Page-rank between minorities and majorities is the same, much more research needs to be conducted to understand how this is reflected in the context of extremal behavior; in the sublinear regime, analytic understanding of the Page-rank distribution in the large network limit is completely open. Further research also needs to be undertaken in the setting where the out-degree distribution depends in a complex manner on the attribute, including settings of heavy tailed out-degree distribution. Similarly this paper has only considered the setting where one has a discrete finite attribute space. The continuous attribute type space will need significantly new techniques.
3. As future 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 (infer) it for the rest of the network considering other samplings methods (e.g. Ribeiro and Towsley 2010).

Acknowledgements Sayan Banerjee is partially supported by the NSF CAREER award DMS-2141621. Shankar Bhamidi and Vladas Pipiras are partially supported by NSF grant DMS-2113662. Sayan Banerjee, Shankar Bhamidi and Vladas Pipiras are partially supported by NSF RTG grant DMS-2134107. We thank two referees for valuable suggestions that significantly improved the presentation and content of the paper.

Author contributions All the authors contributed equally to manuscript.

Data availability https://github.com/gesiscss/Homophilic_Directed_ScaleFree_Networks

Declarations

Conflict of interest The authors declare no competing interests.

References

Antunes N, Bhamidi S, Guo T, Pipiras V, Wang B (2021a) Sampling based estimation of in-degree distribution for directed complex networks. J Comput Graph Stat 30(4):863–876

Antunes N, Guo T, Pipiras V (2021b) Sampling methods and estimation of triangle count distributions in large networks. Netw Sci 9(S1):S134–S156

Antunes N, Banerjee S, Bhamidi S, Pipiras V (2023a) Attribute network models, stochastic approximation, and network sampling and ranking. *Preprint* arXiv:2304.08565v1

Antunes N, Banerjee S, Bhamidi S, Pipiras V (2023b) Learning attribute and homophily measures through random walks. Appl Netw Sci 8(1):39

Antunes N, Banerjee S, Bhamidi S, Pipiras V (2024) Minority representation and relative ranking in sampling attributed networks. In: Cherifi H, Rocha LM, Cherifi C, Donduran M (eds) Complex networks & their applications XII. Springer, Switzerland, pp 137–149

Aral S, Dhillon PS (2018) Social influence maximization under empirical influence models. Nat Hum Behav 2(6):375–382

Aral S, Walker D (2012) Identifying influential and susceptible members of social networks. Science 337(6092):337–341

Banerjee S, Bhamidi S (2021) Persistence of hubs in growing random networks. Probab Theory Relat Fields 180(3–4):891–953

Banerjee S, Huang X (2023) Degree centrality and root finding in growing random networks. Electron J Probab 28:1–39

Barabási A-L, Albert R (1999) Emergence of scaling in random networks. Science 286(5439):509–512

Caliò A, Tagarelli A (2021) Attribute based diversification of seeds for targeted influence maximization. Inf Sci 546:1273–1305

Chebolu P, Melsted P (2008) Pagerank and the random surfer model. In: ACM-SIAM symposium on discrete algorithms, vol 8, pp 1010–1018

Csardi G, Nepusz T (2006) The igraph software package for complex network research. Int J Complex Syst 1695:1–9

Espín-Noboa L, Karimi F, Ribeiro B, Lerman K, Wagner C (2021) Explaining classification performance and bias via network structure and sampling technique. Appl Netw Sci 6(1):78. https://doi.org/10.1007/s41109-021-00394-3

Espín-Noboa L, Wagner C, Strohmaier M, Karimi F (2022) Inequality and inequity in network-based ranking and recommendation algorithms. Sci Rep 12(1):1–14

Galashin P (2016) Existence of a persistent hub in the convex preferential attachment model. Probab Math Stat 36(1):59–74

Garavaglia A, van der Hofstad R, Litvak N (2020) Local weak convergence for pagerank. Ann Appl Probab 30(1):40–79

Granovetter M (1978) Threshold models of collective behavior. Am J Sociol 83(6):1420–1443

Jordan J (2013) Geometric preferential attachment in non-uniform metric spaces. Electron J Probab 18(8):1–15

Karimi F, Génois M, Wagner C, Singer P, Strohmaier M (2018) Homophily influences ranking of minorities in social networks. Sci Rep 8(1):1–12

Kempe D, Kleinberg J, Tardos É (2003) Maximizing the spread of influence through a social network. In: Proceedings of the ninth ACM SIGKDD international conference on Knowledge Discovery and Data Mining, pp 137–146

Lee E, Karimi F, Wagner C, Jo H-H, Strohmaier M, Galesic M (2019) Homophily and minority-group size explain perception biases in social networks. Nat Hum Behav 3(10):1078–1087

Leskovec J, Faloutsos C (2006) Sampling from large graphs. In: Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining, pp 631–636

McPherson M, Smith-Lovin L, Cook JM (2001) Birds of a feather: Homophily in social networks. Ann Rev Sociol 27(1):415–444

Merli MG, Verdery A, Mouw T, Li J (2016) Sampling migrants from their social networks: The demography and social organization of Chinese migrants in Dar es Salaam, Tanzania. Migr Stud 4(2):182–214

Mislove A, Viswanath B, Gummadi KP, Druschel P (2010) 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

Mouw T, Verdery AM (2012) Network sampling with memory: a proposal for more efficient sampling from social networks. Sociol Methodol 42(1):206–256

Page L, Brin S, Motwani R, Winograd T (1999) The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab

Park J, Barabási A-L (2007) Distribution of node characteristics in complex networks. Proc Natl Acad Sci 104(46):17916–17920

Ribeiro B, Towsley D (2010) Estimating and sampling graphs with multidimensional random walks. In: Proceedings of the 10th ACM SIGCOMM conference on internet measurement, pp 390–403

Shrum W, Cheek NH Jr, MacD S (1988) Friendship in school: gender and racial homophily. Sociol Educ 61(4):227–239

Stolte A, Nagy GA, Zhan C, Mouw T, Merli MG (2022) The impact of two types of COVID-19-related discrimination and contemporaneous stressors on Chinese immigrants in the US South. Soc Sci Med Mental Health 2:100159

Sziklai BR, Lengyel B (2022) Finding early adopters of innovation in social networks. Soc Netw Anal Min 13(1):4

Sziklai BR, Lengyel B (2024) Audience selection for maximizing social influence. Netw Sci 12(1):65–87

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, pp 1181–1190