Beyond Assortativity: Proclivity Index for Attributed Networks (PRONE)

Reihaneh Rabbany, Dhivya Eswaran, Artur W. Dubrawski, and Christos Faloutsos

School of Computer Science, Carnegie Mellon University, Pittsburgh PA, USA
{rrabbany,deswaran,awd,christos}@andrew.cmu.edu

Abstract. If Alice is majoring in Computer Science, can we guess the major of her friend Bob? Even harder, can we determine Bob's age or sexual orientation? Attributed graphs are ubiquitous, occurring in a wide variety of domains; yet there is limited literature on the study of the interplay between the attributes associated to nodes and edges connecting them. Our work bridges this gap by addressing the following questions: Given the network structure, (i) which attributes and (ii) which pairs of attributes show correlation? Prior work has focused on the first part, under the name of *assortativity* (closely related to *homophily*). In this paper, we propose PRONE, the first measure to handle pairs of attributes (e.g., major and age). The proposed PRONE is (a) *thorough*, handling both homophily and heterophily (b) *general*, quantifying correlation of a single attribute or a pair of attributes (c) *consistent*, yielding a zero score in the absence of any structural correlation. Furthermore, PRONE can be computed fast in time linear in the network size and is highly useful, with applications in data imputation, marketing, personalization and privacy protection.

Keywords: Attributed Networks, Homophily, Heterophily, Assortativity

1 Introduction

Suppose we know that Alice is majoring in Computer Science. To what extent can we comment on the major of her friend Bob? How accurately can we predict his age or sexual orientation? At a broader level, given the structure of a network and some attributes (e.g., major, age) on the nodes, how can we find out (a) which attributes (b) which pairs of attributes show correlation?

Attributed networks are ubiquitous, occurring in a number of domains. For instance in social networks, where nodes represent people, and edges indicate friendships, the attributes may include interests/demographics of individuals. Similarly in citation networks, where papers (nodes) cite each other (edges), each paper also incorporates information regarding the venue or keywords (attributes). However, despite the prevalence of attributed graphs, the vast majority of network science has dealt solely with the graph structure/topology [5,4] ignoring the attributes.

Studies focusing on the interplay between the network structure and attributes are fairly recent [9,14,15]. For example, in a typical social network, the similarity of individuals motivates them to form relations (social selection) and in turn the individuals

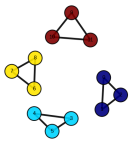
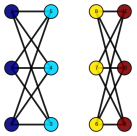
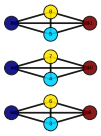
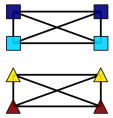
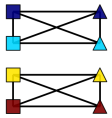
	(i) self-proclivity Given: node's color; Guess: neighbor's colors			(ii) cross-proclivity Given: node's color; Guess: neighbor's shape	
	i.1	i.2	i.3	ii.1	ii.2
					
	homophily	heterophily	random	correlation	no correlation
Q [16]	0.75	-0.25	-0.25	✗	✗
r [17]	**1.0**	-0.33	-0.33	✗	✗
PRONE$_l$	**1.0**	**1.0**	0.21	0.67	**0.0**
PRONE$_2$	**1.0**	**1.0**	0.11	0.5	**0.0**
PRONE$_3$	**1.0**	**1.0**	0.05	0.33	**0.0**

Table 1: All variants of PRONE are **thorough**, **general** and **consistent** in contrast to the baseline *assortativity* measures.

may themselves be affected by their relations (a.k.a. social influence) [14]. This *assortative mixing* and peer influence results in a homophily pattern observed in many real world networks [17], where neighboring nodes exhibit similar characteristics/attributes. Several works use this observation to cluster data [6], build realistic generative models [2,12] and accurate prediction models [1,11]. There are still fewer studies that try to understand assortativity in networks: by quantifying the correlation of nodal attributes and the structure in a static network [17,18], or by investigating the interplay of social selection and influence over time [9].

Assortativity, as a measure for structural correlation of a single attribute, presents a major drawback that it can capture homophily mixing pattern (*i.e.*, when nodes of same attribute value link together) only. This is demonstrated in Table 1. Assortativity (r-index) gives a full score of 1 to perfect homophily (i.1); but is unable to distinguish between perfect heterophily (i.2) and randomness (i.3). Further, it cannot characterize or distinguish the mixing patterns involving a pair of attributes (e.g., ii.1 where there is correlation between color and shape based on structure, and ii.2. where shape and color are independent).

The goal of this work is the formal characterization of the *proclivity* of attributed networks, *i.e.*, *the inclination or predisposition of nodes with a certain value for an attribute to connect to nodes with a certain other value for the same (self-proclivity) or a different attribute (cross-proclivity)*. The problem we address in this work can be informally stated as:

Informal Problem 1 ***Given:*** *an attributed network $\mathcal{G}$, two different attributes a_1, a_2*
To measure:

- *Self-proclivity which captures how predictable neighbors' attribute values for a_1 are, given a node's value for a_1.*
- *Cross-proclivity which captures how predictable neighbors' attribute values for a_2 are, given a node's value for a_1 or vice versa.*

We propose PRONE (PROclivity index for attributed NEtworks) for quantifying both self- and cross- *proclivity* in attributed networks, by drawing upon the clustering validation literature. In place of the confusion matrix (a.k.a. contingency table) which is used to measure the agreements between two groupings of datapoints, we propose to consider the *mixing matrix* (which will be introduced in Section 3) of attributes. PRONE has the following desirable properties:

✓ **Thoroughness**: ability to capture homophily and heterophily
✓ **Generality**: applicability in characterizing both self- and cross- proclivity
✓ **Consistency**: quantification of the absence of correlation as zero
✓ **Scalability**: linear running time with respect to the number of edges

PRONE will help with numerous settings, including:

- *data imputation:* what attributes should we use to guess a missing attribute of Alice, given the attributes of her friends
- *marketing:* for ad placement and enhancing e-shopping experience
- *personalization:* for early depression-detection from online networks [7,8]
- *anonymization/privacy:* which attributes, or pairs of attributes can reveal sensitive information about Alice and thus should be masked

The outline of the paper is as follows. In Section 2, we review related work and present the assortativity indices proposed in literature. Section 3 formally introduces our proposed metric PRONE and Section 4 establishes its theoretical properties. After presenting the results upon applying PRONE to Facebook attribute networks in Section 5, we finally conclude in Section 6.

2 Related Work and Background

In this section, we briefly review the prior work for attributed graphs and present more background on the two assortativity measures proposed in the literature which we will use as our baseline for comparison.

2.1 Related Work

We will group related work under the following four categories: (i) measures for attribute correlation [16,17,18] (ii) dynamic patterns in attributed graphs [9,12] (iii) models for attributed networks [13,19] (iv) link prediction and inference [21,24,14,11,10].

The correlation of attributes with the structure of the network was first studied in [17], in which the assortative mixing of a single attribute was quantified through r-index. To the same end, Q-modularity is proposed [16] based on the surprise in encountering edges connecting attributes of the same value. For vector attributes, assortativity is extended by considering average similarities of connected nodes (*e.g.*, using euclidean or cosine similarity) [18]. There is little work beyond this on quantifying structural correlation of attributes.

On the other hand, several studies try to better understand the dynamics of homophily [9,12]. For example, a clear feedback effect between social influence and selection in the network of Wikipedia editors has been discovered in [9], where they observe

a sharp increase in the average cosine similarity of users right before they interact for the first time followed by a steady increase in their similarity. In a related study, patterns of attributes in Google+ network have been investigated [12] by modeling it as a social-attribute network (SAN), which simply augments the graph by adding nodes which correspond to attribute values and connects them to the individuals who have those attributes. Multiplicative attributes graph model[13] is proposed for attributed networks using a link-affinity matrix, where they assume that the attributes are binary and are independent. To incorporate the attribute correlations into this model, [19] an accept-reject sampling framework was used to filter the edges generated from the underlying model and selectively accept those that match the desired correlations.

Since nodal similarities and social interactions are two tangled factors which affect the evolution of networks [9], models which incorporate the correlation between attributes and relations better predict links and infer attributes, as confirmed by many recent studies [10,11,14,24]. A large body of predictive models extract topological features from the network and combine them with the nodal features to achieve better classification [23] while others directly utilize the generative graph models to jointly predict links and infer attributes [10,11].

We are interested in the more fundamental question of quantifying structural correlations of a single attribute (more general than assortative mixing) or a pair of attributes and thus our work falls into group (i). We will review our only competitors – r-index [17] and Q-modularity [16] in the following section.

2.2 Background

***r*-index** Given an attributed network, r-index for assortativity constructs the $k \times k$ normalized mixing matrix E whose $(i,j)^{th}$ entry, e_{ij}, determines the fraction of edges connecting nodes with attribute value i to nodes with value j. This matrix can be then summarized by an assortativity coefficient [17] defined as:

$$r = \frac{\sum_i e_{ii} - \sum_i e_{i.}e_{.i}}{1 - \sum_i e_{i.}e_{.i}} = \frac{Tr[e] - ||e^2||}{1 - ||e^2||} \tag{1}$$

where $e_{i.} = \sum_j e_{ij}$, $e_{.i} = \sum_j e_{ji}$, and $\sum_{ij} e_{ij} = 1$. Here, $r = 1$ shows perfect assortative mixing and $r = 0$ when there is no assortative mixing.

Q-modularity An alternate characterization of assortativity is to measure how unexpected the edges between the nodes with the same attribute value are compared to *random*. Here, *random* refers to the distribution of edges at random after fixing the degree distribution of the nodes. Mathematically,

$$Q = \sum_i e_{ii} - e_{i.}^2 = \text{Tr}[e] - ||e^2|| \tag{2}$$

Observation: We can see that Q-modularity (Equation 2) is equivalent to the numerator of r-index (Equation 1). In fact, the normalized Q proposed for measuring the assortativity in [16] is equivalent to Equation 1 (since the maximum value of Tr$[e]$ is 1).

3 Proposed Method: PRONE

Consider the $k \times r$ mixing matrix E for two categorical/nominal attributes a_1 and a_2, with respectively k and r distinct values (cardinality). More precisely, elements of E denote *the number of* edges connecting nodes with the corresponding attributes, *i.e.*, e_{ij} represents the number of edges that connect a node that possesses i^{th} value of a_1 ($v_i^{a_1}$) to a node that has the j^{th} value of attribute a_2 ($v_j^{a_2}$). The resulting mixing matrix (and its marginals) is summarized in the following table and form the basis of our PRONE index for measuring the structural correlation between a_1 and a_2.

	$v_1^{a_2}$	$v_2^{a_2}$	...	$v_r^{a_2}$	marginal sums
$v_1^{a_1}$	e_{11}	e_{12}	...	e_{1r}	$e_{1.}$
$v_2^{a_1}$	e_{21}	e_{22}	...	e_{2r}	$e_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$v_k^{a_1}$	e_{k1}	e_{k2}	...	e_{kr}	$e_{k.}$
marginal sums	$e_{.1}$	$e_{.2}$	...	$e_{.r}$	$e_{..}$

Table 2: Mixing matrix of two categorical attributes, a_1 and a_2

Here, we have $e_{i.} = \sum_j e_{ij}$, $e_{.i} = \sum_j e_{ji}$, and $e_{..} = \sum_i \sum_j e_{ij}$. This mixing matrix is analogous to the confusion matrix or contingency table of two clusterings if we assume distinct values of each attribute are class labels for a grouping based on that attribute. Hence, we can quantify the divergence in this matrix as [20]:

$$D_f = \frac{\sum_i [f(e_{i.}) - \sum_j f(e_{ij})] + \sum_j [f(e_{.j}) - \sum_i f(e_{ij})]}{\sum_i f(e_{i.}) + \sum_j f(e_{.j}) - 2\sum_i \sum_j f(\frac{e_{.j}e_{i.}}{e_{..}})} \tag{3}$$

The numerator aggregates the per-row and per-column divergences of this matrix, while the denominator normalizes this quantity using the maximum divergence value when the marginals are fixed. The correlation or agreement of the two attributes a_1 and a_2 is then obtained from $1 - D_f$. We consider three specific derivations of this measure using $f(x) = x \log x$, $f(x) = x^2$, $f(x) = x^3$; the first two correspond to the two most commonly used clustering agreement indexes: respectively Normalized Mutual Information (NMI), and Adjusted Rand Index (ARI). Specifically, if we normalize E so that $e_{..} = 1$, the PRONE$_l$ ($f(x) = x \log x$) and PRONE$_2$ ($f(x) = x^2$) derivations are simplified as:

$$\text{PRONE}_l = \frac{\sum_j e_{.j} \log(e_{.j}) + \sum_i e_{i.} \log(e_{i.}) - \sum_{ij} e_{ij} \log(e_{ij})}{\frac{1}{2}[\sum_j e_{.j} \log(e_{.j}) + \sum_i e_{i.} \log(e_{i.})]} \tag{4}$$

$$\text{PRONE}_2 = \frac{\sum_{ij} e_{ij}^2 - (\sum_i e_{i.}^2)(\sum_j e_{.j}^2)}{\frac{1}{2}[\sum_i e_{i.}^2 + \sum_j e_{.j}^2] - (\sum_i e_{i.}^2)(\sum_j e_{.j}^2)} \tag{5}$$

4 Theoretical Properties

4.1 Thoroughness

PRONE considers all combinations of attribute values when measuring proclivity. Therefore, it can capture all proclivity patterns inherent in the data including homophily and heterophily; whereas the original assortativity index only considers the matched attribute values abd hence can only capture homophily. In particular, PRONE can capture any mixing patterns between the nodes which are regularly link together, *i.e.*, of two given but not necessarily the same attributes.

For instance, in the test case of (i.2) in Table 1, PRONE detects perfect proclivity as red nodes always connect to yellow nodes, and light blue nodes always link to dark blue nodes. This is not captured by the assortativity index (Q or its normalized version r) which only measures the links between nodes of the same color (homophily) and neglects the off-diagonal elements in the mixing matrix E. These indices in fact have the exact same value for (i.2) and (i.3), even though (i.3) has random color assignments and hence zero proclivity. PRONE, however, returns the maximum value 1 for the perfect proclivity in (i.2) and is close to zero for the random case.

Lemma 1 (r-index is not thorough). *R-index does not capture perfect heterophily, especially when the number of attribute values is high.*

Proof. We prove this by giving a counter example. Consider a graph with a single attribute which takes values $\{1, 2, \ldots, 2k\}$ and shows perfect heterophily in the following manner: Nodes with value i (for $i = 1, \ldots, k$) are connected only to nodes with value $k + i$. If p_i is the fraction of total edges connecting nodes of attribute value i with nodes of attribute value $k + i$, the $k \times k$ normalized mixing matrix E is given by $e_{i,k+i} = e_{k+i,i} = p_i/2$ and $e_{ij} = 0$ otherwise. Note that the leading diagonal elements are zero and the row/column sums are $e_{i.} = e_{.i} = e_{k+i.} = e_{.k+i} = p_i/2$ for $i = 1, \ldots, k$. Using these, the r-index (Equation 1) may be calculated as

$$r = \frac{0 - 0.5 \sum_{i=1}^{k} p_i^2}{1 - 0.5 \sum_{i=1}^{k} p_i^2}$$

Taking $p_i = 1/k$, $r = \frac{-1}{2k-1}$. The maximum negative assortativity of -1 is attained only when $k = 2$. As k is increased, the value approaches zero (randomness). Thus r-index fails to capture *perfect heterophily*, particularly for large k. ■

Lemma 2 (Heterophily and self-proclivity). *Perfect heterophily leads to a perfect self-proclivity score of 1, for any choice of f.*

Proof. Let an attribute assume values $1, \ldots, k$ and let π be a permutation of the values such that $\pi_i \neq i$ and $\pi_i = j \iff \pi_j = i$. Let the probability of edge between i and j be $p_i = p_j$ if $j = \pi_i$ and 0 otherwise. Also, let $\sum_i p_i = 1$.

The row/column marginals are $e_{i.} = e_{.i} = p_i$ while $\sum_i \sum_j f(e_{ij}) = \sum_i f(e_{i\pi_i}) = \sum_i f(p_i)$. From Equation 3,

$$\text{PRONE} = 1 - \frac{\sum_i f(p_i) + \sum_j f(p_i) - 2\sum_i f(p_i)}{\sum_i f(p_i) + \sum_j f(p_i) - 2\sum_i \sum_j f(p_i p_j)} = 1$$

■

4.2 Generality

Equation 3 and its PRONE derivations including Equation 4 and Equation 5 do not impose any assumptions on the mixing matrix $E_{r\times k}$ and hence can be applied to general cases. On the other hand, the definition of previous measures for assortativity in Equation 1 and Equation 2 require E to be a square matrix ($r = k$) and hence cannot be extended to measure cross-proclivity of two attributes which have different cardinalities.

For instance, in the test case of (ii.1) in Table 1, we see a mixing pattern between color and shape: *i.e.*, red and yellow circles mix together while light and dark blue squares link to each other. The assortativity measure Q and its normalized version, r, cannot be applied in this case, as the diagonal is not defined for the 4×2 mixing matrix. PRONE, on the other hand, is able quantify this non-square mixing matrix, since it is defined based on average divergence/dispersion in the rows and columns of E. We can see that all variations of PRONE correctly detect a high correlation between shape and color for this case, whereas they return the baseline of 0.0 for the random case of (ii.2) where there is no such correlation.

Lemma 3 (r-index and PRONE). *Squashing the off-diagonal elements in formula of* PRONE$_x$ *yields r-index.*

Proof. Let E be the normalized mixing matrix with $e_{..} = \sum_i e_{i.} = \sum_j e_{.j} = 1$. Using $f(x) = x$, we have

$$\text{PRONE}_x = 1 - \frac{\sum_i e_{i.} + \sum_j e_{.j} - 2\sum_i \sum_j e_{ij}}{\sum_i e_{i.} + \sum_j e_{.j} - 2\sum_i \sum_j e_{i.}e_{.j}} = 1 - \frac{1 - \sum_i \sum_j e_{ij}}{1 - \sum_i \sum_j e_{i.}e_{.j}}$$

Squashing the off-diagonal products to 0 using the indicator function $\mathbb{I}(i = j)$, we get

$$1 - \frac{1 - \sum_i \sum_j \mathbb{I}(i = j)e_{ij}}{1 - \sum_i \sum_j \mathbb{I}(i = j)e_{i.}e_{.j}} = 1 - \frac{1 - \text{Tr}[e]}{1 - \sum_i e_{i.}e_{.i}}$$

which is the expression for r-index. ■

4.3 Consistency

PRONE is expected to return zero when there is no structural correlation in the network. This is a known desired property for the clustering validation indexes. ARI, in particular, is called Adjusted Rand Index for the very same reason that it returns a constant baseline of zero for agreements by chance. This complies with the ~ 0 correlations we observed for random color assignments in the two test cases of (i.3) and (ii.2) of Table 1.

Lemma 4 (Consistency of PRONE). *For any choice of f,* PRONE *is consistent (adjusted for chance),* i.e.*, if values for a nodal attribute are drawn from a categorical distribution ignoring the network structure, its self-proclivity is zero in expectation.*

Proof. Let the multinomial distribution from which the values for attributes a_1 and a_2 are drawn be parameterized by $p_1, \ldots, p_k$ and $q_1, \ldots, q_r$ where k and r are the cardinalities of categorical attributes a_1 and a_2 respectively. Here, $\sum_i p_i = \sum_j q_j = 1$. In the absence of structural correlation of attributes, the expected fraction of edges that connect nodes of attribute values $a_1 = i$ and $a_2 = j$ is $p_i q_j$, which is the expected entry e_{ij} in the normalized mixing matrix E. The expected marginal of row i (or column j) in E is $\sum_j p_i q_j = p_i$ (or q_j). Thus, in expectation,

$$\textsc{ProNE}_f = 1 - \frac{\sum_i f(p_i) + \sum_j f(q_j) - 2\sum_i \sum_j f(p_i q_j)}{\sum_i f(p_i) + \sum_j f(p_j) - 2\sum_i \sum_j f(p_i q_j)} = 0$$

which proves the consistency of ProNE. ■

4.4 Scalability

ProNE has the same computational complexity as the previous measures Q and r, which is the cost of building the mixing matrix E. E can be computed by a single pass over all edges in the graph and hence ProNE is linear in order of number of edges.

In more detail, if we assume m is the total number of edges in the network and k represents the maximum cardinality of attributes, ProNE can be computed in $O(m + k^2)$ time. This matches the computational order for the previous measures, $O(m + k)$, as $k << m$ (the number of edges in a graph is typically much larger than the cardinality of a nodal attribute).

Here, we also empirically measure the computation time of ProNE for networks of varying sizes to show the scalability of the ProNE. In particular, we a generate network of size m, and assign nodes a single attribute with cardinality k, *i.e.*, we assign to each node u, a value in $\{1, \ldots, k\}$ chosen uniformly at random. Figure 1 plots the computational time in seconds as the number of edges grows. The observed linear trend confirms our claim.

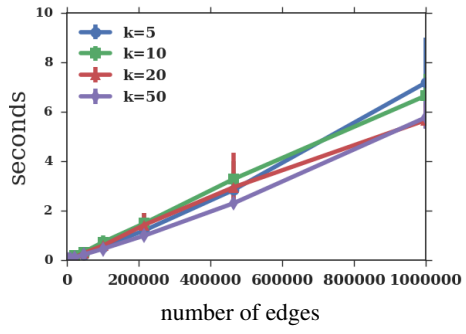


Fig. 1: Scalability of ProNE on networks generated using Barabási and Albert [3] model with $1K$ nodes and $\sim$ $10K$ edges. The attribute cardinality was varied in $\{5, 10, 20, 100\}$ and the results were averaged over 10 runs.

Fig. 2: An example Facebook friendship network, where nodes are colored based on their corresponding attribute value (missing values are white, and non-frequent values are gray). For attribute *status* and *year*, we visually observe some correlation between the color of the nodes and their locations, whereas the locations are derived from a layout algorithm that looks only at the connectivity between the nodes.

Choice of f in practice

Although the above properties are valid for arbitrary choice of f, we recommend choosing f to be a superadditive function[1] satisfying $f(x) \geq 0 \forall x \in [0, 1]$ and $f(1) = 1$ for the proclivity scores to be bounded in $[0, 1]$ [20].

5 Empirical Studies using Real World Data

Here, we study the PRONE in Facebook friendship network of 100 US collages available in a.k.a. *Facebook 100 dataset* [22]. In networks of this dataset, each user has six categorical attributes: (1) *gender* (male/female), (2) *status* (faculty/student/etc), (3) *major*, (4) *second major/minor* (high missing values), (4) *dormitory* of residence, (5) class *year* and (6) *high school*. Figure 2 shows one sample network of this dataset which has 6386 nodes and 217662 friendships edges. The same network is plotted with six different color codings of the nodes, *i.e.*, one plot per attribute in which nodes are colored based on their value for that particular attribute.

In Figure 2, locations of nodes are derived from a network visualization algorithm which only looks at the topology or structure of the graph and tries to place nodes together as cohesive groups. Depending on the layout algorithm used, we can visually observe some of the correlations between attributes (colors) and the structure. In particular, with this example layout, the self proclivity of *year* might be obvious. PRONE provides a fast and quantitative way to detect both the obvious and the hidden structural correlations in such a dataset.

We can see the values of PRONE for Facebook dataset in Figure 2 reported in Table 3. The diagonal of this matrix show the self-proclivity values for the corresponding attributes, and the off-diagonal values provide the cross-proclivity measurements between the corresponding pairs of attributes.

Table 3 reports the results using PRONE$_2$; we observe a similar trend using the PRONE$_l$ and PRONE$_3$ variations. These are reported in Table 4. The choice of PRONE, *i.e.*, the generative function used in Equation 3, depends on the application at hand.

[1] f is superadditive $\iff f(x + y) \geq f(x) + f(y)$

PRONE$_2$	major	gender	year	status	dorm	highschool	minor
major	0.01	0.00	0.00	0.00	0.00	0.00	0.00
gender	0.00	0.00	-0.00	-0.00	-0.00	0.00	-0.00
year	0.00	-0.00	**0.22**	**0.03**	**0.04**	0.00	0.00
status	0.00	-0.00	**0.03**	**0.27**	**0.02**	0.00	0.00
dorm	0.00	-0.00	**0.04**	**0.02**	**0.11**	0.00	0.00
highschool	0.00	0.00	0.00	0.00	0.00	0.00	0.00
minor	0.00	-0.00	0.00	0.00	0.00	0.00	0.00

Table 3: Proclivity of attributes for the Facebook dataset in Figure 2 using PRONE$_2$. The diagonal and off-diagonal entries represent the self-proclivity and the cross-proclivity values respectively. Nodes with missing values were removed before the computation.

	PRONE$_l$							PRONE$_3$						
	major	gender	year	status	dorm	highschool	minor	major	gender	year	status	dorm	highschool	minor
major	0.01	0.00	0.01	0.00	0.01	**0.05**	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
gender	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
year	0.01	0.00	**0.25**	**0.07**	**0.07**	**0.07**	0.01	0.00	0.00	**0.15**	0.01	0.02	0.00	0.00
status	0.00	0.00	**0.07**	**0.09**	0.02	0.02	0.00	0.00	0.00	0.01	**0.29**	0.00	0.00	0.00
dorm	0.01	0.00	**0.07**	0.02	**0.16**	**0.10**	0.02	0.00	0.00	0.02	0.00	**0.05**	0.00	0.00
highschool	**0.05**	0.00	**0.07**	0.02	**0.10**	**0.31**	**0.07**	0.00	0.00	0.00	0.00	0.00	0.00	0.00
minor	0.01	0.00	0.01	0.00	0.02	**0.07**	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 4: Proclivity of attributes for the Facebook dataset in Figure 2 using PRONE$_l$ and PRONE$_3$. These tables provide alternative measurements to Table 3.

We observe similar patterns over different samples in the Facebook 100 dataset. Here, for example, we report the proclivity for another sample, *i.e.*, Rice31 network from this collection which has 4087 nodes and 184828 edges.

Discussion: From the PRONE scores, we infer that the *dormitory* is significantly correlated with friendship as it has a high self-proclivity. This is also the case for *status* (faculty or student) and *year*. What this means is the following: Given Smith's *dormitory* (or *status* or *year*) attribute value, we can predict the *dorm* (or *status* or *year*, respectively) value of his friends. On the other hand, *highschool* and *minor* show zero self-proclivity and the same cannot be said of them. Also, we uncover a surprising pattern that attribute values for *year* and *dorm* show correlation given the friendship network, based on their cross-proclivity of 0.04. Thus, given Smith's *dorm*, it may be possible to predict Smith's friends' *year* values, an inference which is otherwise not possible, from just visualization.

In sum, PRONE is (i) novel and is the first to characterize pairwise attribute correlations given the structure; (ii) is fast to compute and scales linearly with network size; (iii) is effective and discovers interesting correlation patterns when applied to real world graphs. These together make PRONE extremely useful in practice – with applications in anonymizing networks, marketing, data imputation and many more.

	PRONE$_l$							PRONE$_3$						
	major	gender	year	status	dorm	highschool	minor	major	gender	year	status	dorm	highschool	minor
major	0.02	0.00	0.01	0.00	0.01	0.03	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
gender	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
year	0.01	0.00	**0.17**	0.08	0.00	**0.05**	0.01	0.00	0.00	**0.07**	0.02	0.00	0.00	0.00
status	0.00	0.00	0.08	**0.11**	0.00	0.02	0.00	0.00	0.00	0.02	**0.30**	0.00	0.00	0.00
dorm	0.01	0.00	0.00	0.00	**0.25**	**0.09**	0.01	0.00	0.00	0.00	0.00	**0.15**	0.00	0.00
highschool	0.03	0.00	**0.05**	0.02	**0.09**	**0.21**	**0.05**	0.00	0.00	0.00	0.00	0.00	0.00	0.00
minor	0.01	0.00	0.01	0.00	0.01	**0.05**	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00

	PRONE$_2$						
	major	gender	year	status	dorm	highschool	minor
major	0.01	0.00	0.00	0.00	0.00	0.00	0.00
gender	0.00	0.00	0.00	0.00	0.00	0.00	0.00
year	0.00	0.00	**0.13**	**0.05**	0.00	0.00	0.00
status	0.00	0.00	**0.05**	**0.26**	0.00	0.00	0.00
dorm	0.00	0.00	0.00	0.00	**0.24**	0.00	0.00
highschool	0.00	0.00	0.00	0.00	0.00	0.01	0.00
minor	0.00	0.00	0.00	0.00	0.00	0.00	0.01

Table 5: Proclivity of attributes for Rice31 dataset using different derivations of PRONE.

6 Conclusion

In this paper, we proposed PRONE to measure the self- and cross- proclivity patterns and quantify the correlation of a single attribute or a pair of attributes with the network structure. Our proposed PRONE has the following desirable characteristics:

✓ **Thoroughness**: PRONE can capture the full range of mixing patterns in networks, including homophily and heterophily. (Lemma 2)
✓ **Generality**: PRONE can capture both self-proclivity (mixing patterns of a single attribute) and cross-proclivity (mixing patterns of any pair of attributes). (Lemma 3)
✓ **Consistency**: In the absence of structural correlation of nodal attributes, PRONE consistently returns a value of zero in expectation. (Lemma 4)
✓ **Scalability**: PRONE can quantify the mixing patterns, a.k.a. structural correlation, in $\mathcal{O}(m)$ time where m is the number of edges in the network and is fast, processing million-scale graphs in a few seconds.

PRONE is also highly useful, with applications in (i) *data imputation* to guess the values of missing attributes of nodes, (ii) *marketing* for ad-placement, (iii) *personalization* for early depression detection and (iv) *privacy protection* and anonymization of social network.

References

1. Akcora, C.G., Carminati, B., Ferrari, E.: User similarities on social networks. Social Network Analysis and Mining pp. 1–21 (2013)
2. Akoglu, L., Faloutsos, C.: Rtg: a recursive realistic graph generator using random typing. Data Mining and Knowledge Discovery 19(2), 194–209 (2009)
3. Albert, R., Barabási, A.L.: Statistical mechanics of complex networks. Reviews of modern physics 74(1), 47 (2002)
4. Bianconi, G.: Interdisciplinary and physics challenges of network theory. EPL (Europhysics Letters) 111(5), 56001 (2015)

5. Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., Hwang, D.U.: Complex networks: Structure and dynamics. Physics reports 424(4), 175–308 (2006)
6. Bothorel, C., Cruz, J.D., Magnani, M., Micenkova, B.: Clustering attributed graphs: models, measures and methods. Network Science 3(03), 408–444 (2015)
7. Choudhury, M.D., Counts, S., Horvitz, E., Hoff, A.: Characterizing and predicting postpartum depression from shared facebook data. In: CSCW (2014)
8. Colombo, G., Burnap, P., Hodorog, A., Scourfield, J.: Analysing the connectivity and communication of suicidal users on twitter. Computer Communications 73, 291–300 (2016)
9. Crandall, D., Cosley, D., Huttenlocher, D., Kleinberg, J., Suri, S.: Feedback effects between similarity and social influence in online communities. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 160–168 (2008)
10. Gong, N.Z., Talwalkar, A., Mackey, L., Huang, L., Shin, E.C.R., Stefanov, E., Shi, E.R., Song, D.: Joint link prediction and attribute inference using a social-attribute network. ACM Transactions on Intelligent Systems and Technology (TIST) 5(2), 27 (2014)
11. Gong, N.Z., Talwalkar, A., Mackey, L., Huang, L., Shin, E.C.R., Stefanov, E., Song, D., et al.: Jointly predicting links and inferring attributes using a social-attribute network (san). arXiv preprint arXiv:1112.3265 (2011)
12. Gong, N.Z., Xu, W., Huang, L., Mittal, P., Stefanov, E., Sekar, V., Song, D.: Evolution of social-attribute networks: Measurements, modeling, and implications using google+. In: Proceedings of the 2012 ACM Conference on Internet Measurement Conference. pp. 131–144 (2012)
13. Kim, M., Leskovec, J.: Multiplicative attribute graph model of real-world networks. In: Algorithms and Models for the Web-Graph, pp. 62–73. Springer (2010)
14. La Fond, T., Neville, J.: Randomization tests for distinguishing social influence and homophily effects. In: Proceedings of the 19th international conference on World wide web. pp. 601–610. ACM (2010)
15. Lewis, K., Gonzalez, M., Kaufman, J.: Social selection and peer influence in an online social network. Proceedings of the National Academy of Sciences 109(1), 68–72 (2012)
16. Newman, M.: Networks: An Introduction. Oxford University Press, Inc. (2010)
17. Newman, M.E.: Mixing patterns in networks. Physical Review E 67(2), 026126 (2003)
18. Pelechrinis, K., Wei, D.: Va-index: Quantifying assortativity patterns in networks with multidimensional nodal attributes. PloS one 11(1), e0146188 (2016)
19. Pfeiffer III, J.J., Moreno, S., La Fond, T., Neville, J., Gallagher, B.: Attributed graph models: Modeling network structure with correlated attributes. In: Proceedings of the 23rd international conference on World wide web. pp. 831–842. ACM (2014)
20. Rabbany, R., Zaïane, O.: Generalization of clustering agreements and distances for overlapping clusters and network communities. Data Mining and Knowledge Discovery 29(5), 1458–1485 (2015)
21. Silva, A., Meira Jr, W., Zaki, M.J.: Mining attribute-structure correlated patterns in large attributed graphs. Proceedings of the VLDB Endowment 5(5), 466–477 (2012)
22. Traud, A.L., Mucha, P.J., Porter, M.A.: Social structure of facebook networks. Physica A: Statistical Mechanics and its Applications 391(16), 4165–4180 (2012)
23. Wang, P., Xu, B., Wu, Y., Zhou, X.: Link prediction in social networks: the state-of-the-art. Science China Information Sciences 58(1), 1–38 (2015)
24. Yin, Z., Gupta, M., Weninger, T., Han, J.: Linkrec: a unified framework for link recommendation with user attributes and graph structure. In: Proceedings of the 19th international conference on World wide web. pp. 1211–1212. ACM (2010)