PROCLIVITY PATTERNS IN ATTRIBUTED GRAPHS

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

SIAM Workshop on Network Science 2017
July 13–14 · Pittsburgh, PA, USA

Many real world applications include information on both attributes of individual entities as well as relations between them, while there exists an interplay between these attributes and relations. For example, in a typical social network, the similarity of individuals' characteristics motivates them to form relations, a.k.a. social selection; whereas the characteristics of individuals may be affected by the characteristics of their relations, a.k.a. social influence. We can measure proclivity in networks by quantifying the correlation of nodal attributes and the structure [1]. Here, we are interested in a more fundamental study, to extend the basic statistics defined for graphs and draw parallels for the attributed graphs.

More formally, an attributed graph is denoted by (A, X); where $A_{n\times n}$ is the adjacency matrix and encodes the relationships between the n nodes, and $X_{n\times k}$ is the attributes matrix –each row shows the feature vector of the corresponding node. Degree of a node encodes the number of its neighbors, computed as $k_i = \sum_j A_{ij}$. We can extend this notion to networks with binary attributes to the number of neighbors which share a particular attribute x, i.e. $k_i(x) = \sum_j A_{ij}\delta(X_j, x)$; where $\delta(X_j, x) = 1$ iff node j has attribute x.

Similar to the simple graphs, where the degree distribution is studied and showed to be heavy tail, here we can look at: 1) the degree distributions per attribute, 2) the joint probability distribution of any pair of attributes. Moreover, if we assume $A(x_1, x_2)$ is the induced subgraph (or masked matrix of edges) with endpoints of values (x_1, x_2), i.e., $A(x_1, x_2) = A_{ij}\delta(X_i, x_1)\delta(X_j, x_2)$, then we can study and compare these distributions for the induced subgraph per each pair of attribute values. For example, Figure 1 shows the same general trend in the distribution of the original graph and the three possible induced subgraph.

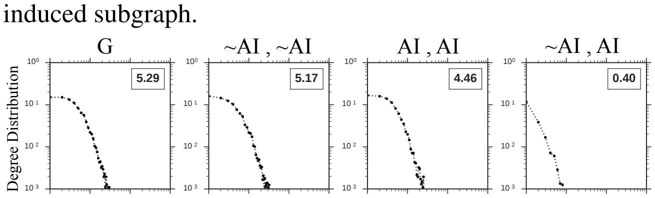

Figure 1: Degree distribution in attribute induced subgraphs for CoRA Citations Network[2], with 11,881 papers with 31,482 citations between them, in which a single attribute AI indicates whether the topic of corresponding paper is in the Artificial Intelligence field.

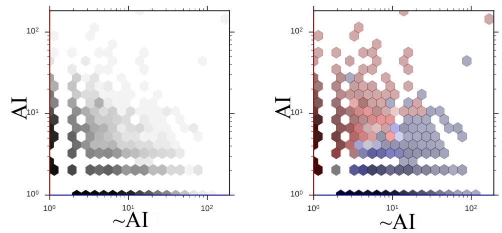

Figure 2: Homophily in degrees: the joint frequency of non-AI neighbors (x-axis), and AI neighbors (y-axis). We can further color divide this distribution based on the value of the node itself, i.e., AI papers are marked by red, and non-AI with blue.

Algebraically, we can compute the degree distributions as the marginals of the adjacency matrix A, i.e., $A\mathbf{1}$. This can be generalized to attributed graphs by considering the matrix multiplication of adjacency matrix A with feature/attribute matrix X, which results in a $n \times k$ matrix AX, in which columns show the degree distribution of nodes for the corresponding attribute value, i.e., number of neighbors of that particular attribute each node has, e.g., number of female friends. In case of two attributes, we can plot the resulted two columns to compare the number of female v.s. male friends per each node. Figure 2 shows such comparison for CoRA, which has a significantly strong proclivity[1] of 0.72, based on the mixing matrix of $\{\{37472, 2380\}, \{2380, 20732\}\}$, due to homophily.

We call AX the "degree matrix", since $(AX)_{ij}$ denotes the number of neighbors that node i has, which have j^{th} attribute, i.e., $(AX)_{ij} = \sum_k A_{ik}X_{kj}$. Here, each column, $(AX)_{:j}$, shows the degree distribution for attribute j, i.e., the number of neighbors which have the j^{th} attribute, per each node; and each row shows the attribute distribution for neighbors of node i, i.e., number of neighbors node i has per each attribute value.

In the same fashion one could study different patterns in attributed networks to reach a better understanding of these ubiquitous datasets which are emerging in diverse domains.

References

[1] R. Rabbany, D. Eswaran, A. W. Dubrawski, and C. Faloutsos. Beyond assortativity: Proclivity index for attributed networks (prone). In *PAKDD*, 2017.

[2] P. Robles-Granda, S. Moreno, and J. Neville. Sampling of attributed networks from hierarchical generative models. In *KDD*, 2016.