Power-law node degree distribution in online affiliation networks

Szymon Chojnacki, Mieczysław Klopotek

Institute of Computer Science PAS, J.K. Ordona 21, 01-237 Warsaw, Poland

sch@ipipan.waw.pl

Abstract. The purpose of this article is to present a preferential attachment model adjusted to generation of bipartite graphs. The original model is able to produce unipartite graphs with node degree distribution following power law relation. The motivation for extending classic model is the fact that multi-partite graph topologies are becoming more and more popular in social networks. This phenomenon raises questions about our ability to transform models that describe unipartite graphs to the new settings. In some cases bipartite structure is given explicitly, in other cases the bipartite structure may be induced from more complex topologies. We present both empirical results concerning node degree distribution in real-life bipartite networks and a modified preferential attachment model that is able to reflect the properties observed in these networks.

1 Introduction

Bipartite or affiliation networks describe a situation in which we have two types of nodes and direct links only between nodes of different kinds. One type of nodes could be interpreted as actors and the second as events in which actors take part [1]. From this network we could induce two unipartite graphs containing only nodes of one type. However, it has been shown that despite some similarities, in general the structural properties of both representations may differ significantly [2].

More formally, a graph is an ordered pair $G = (V, E)$ comprising a set of vertices (or nodes) $V$ and a set of edges (or links) $E$. A bipartite network (or bigraph) is a graph which vertices can be labeled by two types $A$ and $B$. The difference with classic unipartite graph is the fact that $V$ consists of two sets $V = \{V_A \cup V_B : V_A \cap V_B = \emptyset\}$ and edges exist only between nodes of different types $E \subseteq V_A \times V_B$. Degree of a node is the number of its direct neighbors. Probability that a degree $k_v$ of node $v$ of type $A$ is exactly $z$ can be calculated as:

$$P(k_v = z) = \frac{|\{v \in V_A : |\{w : (v, w) \in E\}| = z\}|}{|V_A|}.$$  \hfill (1)

In this article we focus on degree distributions obtained from our graph generative model. Empirical results show that the density function of a degree distribution in real life datasets often follows power law relation. It means that the frequency of an event decreases at a higher rate than its size increases. Such property is obtained from a polynomial relation:
In the above equation $a$ and $k$ are constants and $k$ is called scaling exponent. The tail in power law distribution vanishes slower than exponentially, which is described as a heavy-tail. The shape of density function in scale invariant (or scale free) distribution does not change when we multiply variable $x$ by $c$. One can verify that:

$$f(cx) = a \cdot (cx)^{-k} = c^{-k}f(x) \propto f(x).$$

This property is visible when we take logarithm of both sides of Eq. 2. In this way we obtain a linear relationship between variables $\log(f(x)) = -k \log(x) + \log(a)$ in which $k$ is responsible for the slope of a line drawn by the points.

The scale invariant distribution is unavailable for purely random graphs [3]. The first model that was able to reflect this property by an atomic process was preferential attachment model [4]. We modify this model and show that our extension is also able to create power law degree distribution in bipartite networks.

The rest of the paper is organized as follows: Section 2 surveys the related work. Section 3 gives results of our exploratory analysis. In Section 4 we describe in detail the preferential attachment model for bipartite random graphs generation, we conduct both formal mathematical analysis and show the results of experimental simulations. We conclude and discuss the implications of our findings in the last fifth section.

## 2 Related work

In preferential attachment model we initialize graph with a set of $m_0$ vertices. In each iteration new node is added to a network and is connected to $p$ nodes. A probability that a node already existing in a graph would be connected to a new node is proportional to its degree [4]. This process reflects the rich gets richer phenomenon. It generates asymptotical power law node degree distribution with scaling exponent equal to 3. Extensions of the base model were proposed to build networks with the exponent ranging from 2 to infinity. In winners don’t take all model probability that a node will be selected is modified by a mixture of preferential attachment and uniform distribution [5]. The same idea results from stochastic urn transfer model [6]. Generalization described in [7] allows to add edges between nodes existing in a network during each iteration. The mixture of three processes in one iteration: adding a node, adding an edge and rewiring an edge was proposed in [1]. Copying model is an example of a graph generator that results in power law distribution and is not based on preferential attachment [8, 9].

Results of exploratory analysis conducted on real world networks have lead to additional observations beyond node degree distribution and construction of models that try to reflect these characteristics. Small diameter of real networks resulting from Milgram’s experiments [10], which is usually referred to as small world phenomenon or six degrees of separation is explained by models that allow local structures to be connected to random distant nodes [11, 12]. It has been shown that the proportion of the number of edges to the number of nodes increases over time. This densification feature was explained by a generic Community Guided Attachment model, in which an underlying
hierarchical structure of nodes is considered and a probability of a link between two nodes is inversely proportional to the distance in a tree between the two nodes [13]. It has been shown that an effective diameter of a network decreases over time [13] and a very powerful model to simulate this pattern, based on adjacency matrix multiplication, was proposed [14]. In [9] the users of a social network were divided into three groups: linkers, inviters and passive, which enabled to created a model that simulates the structure of largest connected component and middle region in a real network.

Most of the described research was focused on the analysis of unipartite graphs. Early analysis of affiliation networks was limited by the size of available data. In [15] a network of twenty-six chief executive officers and their membership in fifteen civic clubs and corporate boards was analyzed. Even if row data have bipartite structure, the analysis is usually limited to the projection onto one of the modalities. However, it has been shown that even though some properties are common for both views (e.g. relative size of the largest connected component), other important measures may differ (e.g. clustering coefficient) [2]. In a model for bipartite graph generation described in [2] number of nodes is fixed and we draw degree for each node from a predefined distribution, in the second step the ends of generated edges are connected randomly. An iterative model for bipartite graph generation was proposed in [16]. In each iteration one node of preselected modality is added, its degree is drawn from a predefined distribution and for every edge a decision is made whether to join it to existing nodes following preferential attachment or join it to a newly created node. Empirical results show that node degree distribution in random graphs output from this process resembles that of real-world datasets.

In this article we give empirical evidence for an observation that node degree distribution in affiliation networks induced from online social networks is power law and present modified version of preferential attachment model, which iteratively generates network meeting this criterion.

### 3 Observations

In this section we describe degree distributions obtained for two bipartite large graphs built from three real-life datasets. The datasets are defined in the first subsections. In the second subsection we describe the results.

#### 3.1 BibSonomy dataset

The BibSonomy dataset was used during ECML/PKDD 2009 Discovery Challenge [17]. It contains a full snapshot of BibSonomy.org users’ activity until the end of 2008. The users of BibSonomy(\(\text{http://www.bibsonomy.org}\)) can bookmark favorite publications and web pages and assign tags (keywords) to saved resources. We utilized tas (tag assignments) file which contained 401 104 tag assignments made by 3 617 user for 235 328 resources using 93 756 distinct tags. From this data we built a bipartite graph of user-resource connections.
3.2 CiteULike dataset

The structure of main table in this dataset is analogous to the tag assignment table in the BibSonomy dataset. The two datasets differ in magnitude and type of used resource, in case of CiteULike (http://www.citeulike.org) we had publications instead of web pages. We obtained additional table with assignments of CiteULike users to thematic groups. Tas file contains 2 657 227 tag assignments made by 18 467 users for 557 101 articles by means of 166 504 tags. The second file linking groups with users contains only 2 336 groups and 5 208 users. We built a bipartite graph of user - group connections.

3.3 Degree distributions

The degree distributions of nodes of two types for the eight bipartite graphs are presented at Fig. 1 and Fig. 2. In most cases the points are shaped in a straight line on a log-log scale, which is a necessary condition for power law distribution. We obtained scaling exponent by solving least squares linear regression problem. We checked which observations are overly influential with respect to DFFITS statistic [18]. For most distribution, there are outliers at the beginning and at the end of the domain. Therefore, we considered only points from the fifth to the twenty fifth for cleaned parameters estimation. Estimated exponents are in (1,4) interval.

Fig. 1. Node degree distributions for BibSonomy dataset. A straight line of points on a log-log scale is characteristic for power law distributions. The dotted line is obtained by estimation of linear regression with least squares method for points between the fifth and the twenty fifth (overly influential points are omitted). Scaling coefficients are given next to the dotted lines.
Fig. 2. Node degree distributions for CiteULike dataset.

The presented empirical results indicate that similarly to unipartite graphs, bipartite online graphs exhibit power-law feature of node degree distributions. This property was explained by preferential attachment mechanism in unipartite graphs. In the following section we generalize the mechanism to the setting of bipartite graphs.

4 Generative procedure

Our model can be perceived as an extension of classic preferential attachment model for bipartite graphs. In the following subsections we present the generative procedure and check stability of estimated exponents.

4.1 Detailed description

Initially we start with a small number of edges $m_0$ connecting $m_0$ vertices of type $A$ with $m_0$ vertices of type $B$. During each iteration we add a node $v$ of type $A$ and a node $w$ of type $B$ to the graph. We connect both nodes with $p$ and $q$ nodes of different types respectively (Fig. 3). We call $p$ and $q$ basic node degree.

The selection of the nodes that we will connect to is based on a preferential attachment rule, which states that the probability that a new edge will be connected to a node is proportional to its degree. After $t$ iterations the number of vertices and edges is equal to $|V(t)| = 2 \cdot (t + m_0)$ and $|E(t)| = t \cdot (p + q) + m_0$. If we use $k_i$ to represent the degree of node $i$, then for a bipartite graph Eq. 4 describes the relation between the number of edges and the sums of node degrees.
During each iteration one node of type A and one node of type B are added to the graph. A node of type A is connected to \( p \) nodes of type B and a node of type B is connected to \( q \) nodes of type A. The selection of nodes is random, based on preferential attachment principle (probability of being selected is proportional to nodes degree).

\[
\sum_{i \in V_A} k_i = \sum_{j \in V_B} k_j = |E| \tag{4}
\]

Degree \( k_i \) of node \( i \) of type B changes over time with accordance to the following differential equation:

\[
\frac{\partial k_i}{\partial t} = \frac{k_i \cdot p}{(p + q) \cdot t} \tag{5}
\]

The denominator of right side of the equation is equal to the number of edges in graph \( G(t) \) when \( t >> m_0 \). The nominator is a multiplication of node’s degree times \( p \). If we sum the right side of the equation over all nodes of type B we obtain \( p \), which is the amount of edges that we added to nodes of type B during one iteration. Moreover the right side is also proportional to node degree and therefore satisfies preferential attachment rule.

4.2 Experiments

In this subsection we verify the impact of initial number of edges on the stability of scaling exponents in graphs obtained after finite number of iterations. The results let us suspect that power law distribution is obtained early. We simulated the above described
Fig. 4. Estimated scaling exponent for node degree distribution of node modalities obtained by changing initial number of edges. All points were used for least squares regression estimation. Node’s degree during first attachment was set as $p = 2$ and $q = 3$.

Fig. 5. Node of type B degree distribution and estimated scaling exponents. Graph generated after 20 000 iterations, for initial number of edges $m_0 = 5$ and base degrees $p = 2$ and $q = 3$. We can see a slightly increasing trend of the value of the exponent for both modalities (Fig. 4). It means that if the number of initial edges (and initial nodes with degree equal
to one) is higher than the generated distribution is steeper. In case of nodes of type B, which have higher basic degree, the exponent destabilizes for smaller number of initial edges. We verified the deviant distribution and the reason for flatter distribution shape is the fact that some of the $m_0$ nodes were not selected by preferential attachment rule even once during 20 000 iterations.

Based on visual analysis of generated distributions we asserted that in all situations we obtained a straight line shape of points (except for outliers), which is a good indicator of power law distribution. In Fig. 5 we present the distribution for one of generated graphs.

5 Conclusion

In this article we analysed node degree distributions in bipartite real-world graphs. It turns out that power law relation is common in such networks. This observation motivated us to modify classic preferential attachment model and analyze its properties in case of bigraphs. We showed that preferential attachment mechanism can be used to generate random affiliation networks with power-law node degree distributions. The results of performed simulations suggest that the number of initial edges in random graphs influences the stability of estimated scaling exponent even after several thousand of iterations.

References

Power-law node degree distribution in online affiliation networks


